

**"Combines philosophical investigation and historically informed introduction.
If you want a comprehensive guide to analytic philosophy, look no further."**

Mark Textor, Professor of Philosophy, King's College London, UK

THE BLOOMSBURY COMPANION TO ANALYTIC PHILOSOPHY

EDITED BY
**BARRY DANTON AND
HOWARD ROBINSON**

B L O O M S B U R Y

The Bloomsbury Companion to
Analytic Philosophy

Bloomsbury Companions

The *Bloomsbury Companions* series is a major series of single-volume companions to key research fields in the humanities aimed at postgraduate students, scholars, and libraries. Each companion offers a comprehensive reference resource giving an overview of key topics, research areas, new directions, and a manageable guide to beginning or developing research in the field. A distinctive feature of the series is that each companion provides practical guidance on advanced study and research in the field, including research methods and subject-specific resources.

Titles currently available in the series:

Aesthetics, edited by Anna Christina Ribeiro

Aristotle, edited by Claudia Baracchi

Continental Philosophy, edited by John Mullarkey and Beth Lord

Epistemology, edited by Andrew Cullison

Ethics, edited by Christian Miller

Existentialism, edited by Jack Reynolds, Felicity Joseph and Ashley Woodward

Hegel, edited by Allegra de Laurentiis and Jeffrey Edwards

Heidegger, edited by Francois Raffoul and Eric Sean Nelson

Hobbes, edited by S. A. Lloyd

Hume, edited by Alan Bailey and Dan O'Brien

Kant, edited by Gary Banham, Dennis Schulting and Nigel Hems

Leibniz, edited by Brendan Look

Locke, edited by S.-J. Savonius-Wroth, Paul Schuurman and Jonathan Walmsley

Metaphysics, edited by Neil A. Manson and Robert W. Barnard

Philosophy of Language, edited by Manuel García-Carpintero and Max Kölbel

Philosophy of Mind, edited by James Garvey

Philosophy of Science, edited by Steven French and Juha Saatsi

Plato, edited by Gerald A. Press

Pragmatism, edited by Sami Pihlström

Socrates, edited by John Bussanich and Nicholas D. Smith

Spinoza, edited by Wiep van Bunge

The Bloomsbury Companion to Analytic Philosophy

Edited by

Barry Dainton

and

Howard Robinson

Bloomsbury Academic
An imprint of Bloomsbury Publishing Plc

B L O O M S B U R Y
LONDON • NEW DELHI • NEW YORK • SYDNEY

Bloomsbury Academic
An imprint of Bloomsbury Publishing Plc

50 Bedford Square
London
WC1B 3DP
UK

1385 Broadway
New York
NY 10018
USA

www.bloomsbury.com

Bloomsbury is a registered trade mark of Bloomsbury Publishing Plc

First published 2014

First published in paperback 2015

© Barry Dainton, Howard Robinson and Contributors, 2014, 2015

Barry Dainton and Howard Robinson have asserted their right under the Copyright, Designs and Patents Act, 1988, to be identified as Editors of this work.

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage or retrieval system, without prior permission in writing from the publishers.

No responsibility for loss caused to any individual or organization acting on or refraining from action as a result of the material in this publication can be accepted by Bloomsbury Academic or the author.

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

ISBN: PB: 978-1-4742-3649-2
HB: 978-1-4411-2628-3
ePDF: 978-1-4742-3647-8
ePub: 978-1-4742-3648-5

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress.

Typeset by Newgen Knowledge Works (P) Ltd, Chennai, India

Contents

Acknowledgments	viii
Overview	ix
Preface	x
Contributors	xvi
Part I: History, Methods, and Problems	1
1 A Different World <i>Barry Dainton</i>	3
2 From Idealism to a Realistic (Platonic) Pluralism <i>Barry Dainton</i>	8
3 Principia Ethica <i>Barry Dainton</i>	14
4 <i>Principles</i> and Paradox <i>Barry Dainton</i>	18
5 Frege <i>Barry Dainton</i>	25
6 On Denoting, Acquaintance, and Construction <i>Barry Dainton</i>	31
7 Wittgenstein and the <i>Tractatus</i> <i>Barry Dainton</i>	39
8 The Vienna Circle and its “Wissenschaftliche Weltauffassung” <i>Barry Dainton</i>	50
9 Later Wittgenstein <i>Barry Dainton</i>	61
10 Quine <i>Barry Dainton</i>	71

Contents

11	Oxford and Ordinary Language <i>Barry Dainton</i>	83
12	Developments in Ethics <i>Barry Dainton</i>	99
13	Davidson <i>Barry Dainton</i>	108
14	Kripke and Putnam <i>Barry Dainton</i>	118
15	Analytic Philosophy of Mind <i>Barry Dainton and Howard Robinson</i>	128
	Appendix: A Simple Introduction to Tarski's Theory of Truth <i>Barry Dainton</i>	138
	Notes	148
	Part II: Current Research and Issues	161
	Introduction to Part II <i>Barry Dainton and Howard Robinson</i>	163
16	Mathematics and Logic <i>Mary Leng</i>	172
17	Philosophy and Language <i>Barry C. Smith</i>	201
18	Meaning, Normativity, and Naturalism <i>Richard Gaskin</i>	230
19	Philosophy of Science <i>James Ladyman</i>	255
20	Philosophy of Physics <i>Barry Loewer</i>	285
21	Causation <i>Helen Beebe</i>	312
22	Metaphysics <i>E. J. Lowe</i>	336
23	Philosophy of Mind: Consciousness, Intentionality and Ignorance <i>Daniel Stoljar</i>	355
24	Personal Identity: Are We Ontological Trash? <i>Mark Johnston</i>	378

25	Free Will <i>Ferenc Huoranszki</i>	415
26	Knowledge <i>Bryan Frances and Allan Hazlett</i>	432
27	The Philosophy of Perception: An Introduction <i>Paul Snowdon</i>	453
28	Practical Reasons: The Problem of Gridlock <i>Ruth Chang</i>	474
29	Moral Demands and Ethical Theory: The Case of Consequentialism <i>Attila Tanyi</i>	500
30	Political Obligation, and the Site and Scope of Justice <i>Andres Moles</i>	528
	Part III: New Directions in Analytic Philosophy <i>Barry Dainton and Howard Robinson</i>	551
31	Coda A: What is Analytic Philosophy? <i>Barry Dainton and Howard Robinson</i>	569
32	Coda B: Analytic versus Continental <i>Howard Robinson</i>	575
	Notes	579
	Chronology	582
	Timeline of Individual Philosophers	586
	A–Z of Key Terms and Concepts	589
	Resources	628
	Annotated Bibliography	631
	Bibliography	647
	Author Index	659
	Subject Index	668

Acknowledgments

The editors wish to thank everyone who helped with the project. Particular thanks to Richard Gaskin for his detailed comments on Part I, and Bloomsbury editors Sarah Campbell and Rachel Eisenhauer, for their advice, and for their patience.

Overview

In this Companion we provide a guide to analytic philosophy's past, present, and future; we also attempt to specify what—if anything—is genuinely distinctive about it.

In Part I we provide an historical introduction to the movement, one that takes in Russell and Moore's rejection of absolute idealism at the turn of the twentieth century, Frege's contributions to logic and the foundations of mathematics, some of Russell's subsequent work—including his influential theory of descriptions—Wittgenstein's early and later philosophies, logical positivism, and the "ordinary language" period of Oxford philosophy. We also introduce some of the more influential doctrines of four important American philosophers—Quine, Davidson, Putnam, and Kripke—along with more general developments in analytic ethics, and the philosophy of mind. The historical survey ends with the early 1980s. Although it makes no pretense at being exhaustive, the survey does cover the main episodes in the development of the movement in a manner that—we hope—helps bring what is truly distinctive about analytic philosophy into clear view.

Part II aims to present the state-of-the-art in the major areas of analytic philosophy. Analytic philosophy is a specific movement in philosophy, with a distinct history, but the main burden of the articles is not historical: they exhibit current analytic philosophy in action. The historical material in Part I will lead the reader to the point from which the articles take off.

In Part III we turn to more recent developments, and venture some speculations as to what the near future may hold. Drawing on the earlier historical survey, we also broach the controversial question of what analytical philosophy *is*, and, in particular how it differs from its foil—so-called Continental Philosophy. As further aids to the reader we have included an extensive glossary of key terms and concepts, an annotated bibliography, timelines of major events and publications, and a guide to further resources.

Preface

In his 1931 essay “The Future of Philosophy,” Moritz Schlick—logical positivist and founding member of the Vienna Circle—observed that the history of philosophy hitherto had been one of profound and irreconcilable disagreement. The doctrines of Plato are radically different from those of Aristotle, likewise the metaphysical systems of Leibniz and Spinoza, Kant, and Hegel. This undeniable historical fact might naturally lead one to doubt whether any real progress had been made in Western philosophy in the two millennia of its history; indeed, it might lead one to doubt whether genuine progress in philosophy—the sort of progress that the natural sciences have been enjoying since the seventeenth century—was even possible. Schlick goes on to pose these questions:

Will this chaos that has existed so far continue to exist in the future?
Will philosophers go on contradicting each other, ridiculing each other’s opinions, or will there finally be some kind of universal agreement, a unity of philosophical belief in the world? (1931)

He goes on to note that there is a further consideration that might incline one to pessimism on this score: the fact that many of the competing schools have each had their own different and distinctive *methods* for arriving at philosophical truths. Descartes’ method of doubt is very different from Spinoza’s axiomatic approach, Kant’s Copernican Revolution takes us in a different direction entirely, as does Hegel’s dialectic—and similarly for Heidegger’s proposal that we return to the “question of Being” and attempt to hear its call.

Despite all this, Schlick suggests there are reasons for being optimistic that the long reign of chaos will soon be brought to a close: philosophers will soon stop ridiculing each other’s opinions. Why? Because an entirely *new* philosophical methodology has appeared on the scene. Thanks to this new approach the centuries-long impasse will be ended. What is this revolutionary new method? Instead of attempting to uncover distinctively philosophical

truths about the nature of reality, philosophers will instead devote their attention to *analyzing propositions*, with a view to clarifying their true meaning. Schlick continues:

The view which I am advocating has at the present time been most clearly expressed by Ludwig Wittgenstein; he states his point in these sentences: "The object of philosophy is the logical clarification of thoughts. Philosophy is not a theory but an activity. The result of philosophy is not a number of 'philosophical propositions', but to make propositions clear."¹

This Wittgensteinian doctrine is one important strand of a more general movement or tradition, that of analytic philosophy. This movement is usually seen as having originated in the early years of the twentieth century with the work of the Cambridge philosophers G. E. Moore and Bertrand Russell, and (a few years later) Ludwig Wittgenstein. Over the course of the twentieth century a variety of doctrines and approaches were espoused by analytic philosophers—the doctrines of the "early" and "later" Wittgenstein are notoriously divergent; Russell's views were continually evolving. Even so, the analytic movement is associated with some characteristic features: (a) the placement of the analysis of language (or concepts) at the very center of the philosophical stage, (b) exploiting whenever possible the major breakthroughs in symbolic logic made in the late nineteenth and early twentieth centuries by Frege, Peano, Russell, and others, (c) a passionate commitment to clarity of exposition, care in drawing distinctions, rigor in argument, and (d) the fostering of communal self-criticism. As we will see subsequently, (c) and (d) are more or less ubiquitous, (a) and (b) are not.

During the short history of analytic philosophy doctrinal evolution has gone hand-in-hand with geographical diffusion: first to the Vienna of the 1930s (where Schlick's "Circle" met), then to the Oxford of the 1940s and 1950s, where the "ordinary language" movement flourished, but also (and importantly) to the United States—where most of the Vienna Circle moved, when its members sought to refuge from the rise of Nazism in the European continent in the 1930s. By the 1950s the analytic movement had come to dominate the entirety of the English-speaking world—including Australia, which has long been an influential center in its own right—and it has retained this dominant position to the present day. Indeed, its influence is now rapidly expanding beyond its Anglo-Saxon homelands. Since the 1990s there has been a rapid expansion of interest in the analytic brand of philosophy in other (non-Anglophone) parts of the world, which has led in turn to an explosion of organizations devoted to its promulgation, both in Continental Europe—much of

which was long regarded as hopelessly hostile territory. The manifesto of the European Society for Analytic Philosophy (ESAP) runs thus:

However convenient the opposition between “Analytic” and “Continental” philosophy may be, it is inadequate, for there are analytic philosophers on the Continent, and the values and aspirations of analytic philosophers are (meant to be) universal. Analytic philosophy is characterized above all by the goal of clarity, the insistence on explicit argumentation in philosophy, and the demand that any view expressed be exposed to the rigours of critical evaluation and discussion by peers. The universality of these values is one of the reasons for the current revival of analytic philosophy in continental Europe after the long interruption due to the Second World War and the North American exile of many European philosophers.²

It does not stop in Europe: analytic philosophy has also taken root in Latin America and Asia. Surveying the penetration of analytic philosophy in Latin America, Perez and Ortiz-Milan conclude: “. . . a huge amount of analytic philosophy has been produced [in Latin America] in the last fifteen years or so . . . Analytic philosophy has a promising future in Latin America: no doubt it will keep growing in the years to come” (2010, p. 211). In their survey of analytic philosophy in China, Jiang and Bai summarize thus: “Although its influence on Chinese intellectuals is not as strong as that of continental philosophy (for example, phenomenology), analytic philosophy has been one of the main influential Western philosophies in China since the beginning of the twentieth century.”

While Schlick would no doubt have been pleased by the persistence and flourishing of the analytic movement of which he was an early and prominent proponent, he would surely be disappointed by the continuing disagreements that mark the contemporary philosophical scene.³ Some of these disagreements separate the analytic philosophers from their “continental” opponents. But other disputes—many scarcely less profound—divide analytic philosophers themselves. The current analytic scene contains those who believe that the philosophy of mind is prior to the philosophy of language (contrary to the analytic orthodoxy) and those who believe the opposite. There are also those who hold the mental to be reducible to the physical and those who reject this; there are “theories of meaning” of radically different kinds; those who favor standard two-valued logic are numerous, but so too are those who favor one or other brand of “deviant logic.” Some believe that ethical values are objective, but there are also many who hold that values are subjective or culturally relative; there are many who regard philosophy as being continuous with the natural sciences, and there are also many who believe that philosophy has a

legitimate sphere of its own that is distinct from that of the natural sciences—the list of significant divergences on key issues goes on (and on). Evidently, the “chaos” that Schlick bemoaned has continued; the analytic movement has not (yet) delivered the sort of progress its early pioneers hoped for: progress that is slow, steady, and incremental, but also irreversible and indisputable.

Interestingly—and perhaps importantly—this failure has not, on the whole, led analytic philosophers seriously to doubt that their movement is a special and distinctive one, perhaps even a revolutionary one. Here is Peter Hacker (arguably the preeminent contemporary follower and interpreter of Wittgenstein) assessing the achievements of analytic philosophy:

From the beginning of the century and for the next seventy-five years, it was, in all its various transformations, the most distinctive style of philosophical thought of our times. To the extent that the seventeenth and eighteenth centuries can be characterized as above all the age of reason and enlightenment in philosophy, and the nineteenth as the age of historicism and historical self-consciousness, then to that extent the twentieth century can be said to have been the age of language and logic. The task of exploring the philosophical consequences of the thought that man is above all a language-using creature fell to analytic philosophy. So too did that of clarifying the significance of the unprecedentedly powerful formal logic invented at the turn of the century, and of elucidating the relations between logical calculi, language and thought. (Hacker 1996, p. ix)

John Searle, a leading philosopher of language and mind, goes even further:

Given [analytic philosophy’s] constant demand for rationality, intelligence, clarity, rigour and self-criticism, it is unlikely that it can succeed indefinitely, simply because these demands are too great a cost for many people to pay. The urge to treat philosophy as a discipline that satisfies emotional rather than intellectual needs is always a threat to the insistence on rationality and intelligence. However, in the history of philosophy, I do not believe we have seen anything to equal the history of analytic philosophy for its rigour, clarity, intelligence and, above all, its intellectual content. There is a sense in which it seems to me that we have been living through one of the great eras in philosophy. (2003, p. 21)

Searle’s contention that the analytic school has produced *intellectual content* of an unprecedented quality is without question one with which the vast bulk of contemporary analytic philosophers would concur, despite their other disagreements.

But have the good times come to an end? Has analytic philosophy attained something approaching world-dominance just as the movement has ceased to *be* a movement at all, in any meaningful sense? In the eyes of some—in both analytic and continental circles—this is indeed the case. When Hacker writes that analytic philosophy was the most distinctive mode of philosophical thought “from the beginning of the century and for the next seventy-five years” he is clearly implying that this situation did not extend beyond the mid 1970s. Elsewhere he is more explicit: “Analytic philosophy flourished in various forms from the 1910s until the 1970s. In the last quarter of a century, however, it has lost its distinctive profile” (1998, p. 3). Brian Leiter, in the introduction to his recent anthology *The Future for Philosophy* is more blunt in his assessment:

Philosophy today—especially, though not only, in the English-speaking countries—is not a monolith, but a pluralism of methods and topics. “Analytic” philosophy, for example, the target of many polemics by those with little knowledge of the discipline, is defunct. (2004, p. 1)

Half a decade earlier, in summarizing the state of play in their *The Story of Analytic Philosophy*, Biletzki and Matar venture a verdict that is only slightly less pessimistic:

It seems beyond question that analytic philosophy has been, for some time now, in a state of crisis—dealing with its self-image, its relationships with philosophical alternatives, its fruitfulness and even its legitimacy in the general philosophical community. (1998, p. xi)

Of course, enduring a “crisis” is not quite so serious as dying, but neither is it the same as enjoying a state of blooming health. Searle offers a more optimistic diagnosis:

... it has passed from being a revolutionary minority point of view held in the face of traditionalist objections to becoming itself the conventional, establishment point of view. Analytic philosophy has become not only dominant but intellectually respectable, and like all successful revolutionary movements, it has lost some of its vitality in virtue of its very success. (ibid., p. 21)

So for Searle at least, pronouncements of the death of analytic philosophy are premature; the movement has merely reached maturity, and suffered the inevitable consequences; loss of youthful élan is entirely compatible with enjoying a long and vigorous middle-life, or so many of us hope.

As to which of these verdicts on the current state of health of analytic philosophy is the closer to the truth, we will not venture an opinion here—though we will comment further in Part III. However one thing is already entirely clear. For those with an interest in matters philosophical, the fate and condition of analytic philosophy is an issue of very considerable significance.

Contributors

Helen Beebe is Samuel Hall Professor of Philosophy at the University of Manchester. Much of her research has been concerned with “Humean” themes, and she has published articles on the laws of nature, natural necessity, inductive skepticism, free will, and the observability of causation. On Hume himself she has written *Hume on Causation* (Routledge, 2006), and she has coedited (along with P. Menzies and C. Hitchcock) *The Oxford Handbook of Causation* (OUP 2009).

Ruth Chang is Professor of Philosophy at Rutgers University, New Brunswick. Before receiving her doctorate in philosophy from Oxford, she gained a doctorate in law from Harvard Law School and she has taught in both philosophy and law departments. She is author of *Making Comparisons Count* (Routledge, 2001) and editor of *Incommensurability, Incomparability, and Practical Reason* (Harvard University Press, 1997).

Bryan Frances has been a Professor at Fordham, and prior to that a lecturer at Leeds University. He received an MA in physics, before doing his PhD in philosophy at the University of Minnesota. His main publications are in epistemology and the philosophy of mind and he is author of *Scepticism Comes Alive* (OUP, 2005).

Richard Gaskin is Professor of Philosophy at the University of Liverpool, and has held visiting fellowships at the universities of Edinburgh, Bonn, and Mainz. He has published extensively in the areas of philosophy of language, history of philosophy, philosophy of literature, and the classical tradition. His most recent books are *Experience and the World’s Own Language: a Critique of John McDowell’s Empiricism* (OUP, 2006); *The Unity of the Proposition* (OUP, 2008), *Language, Truth, and Literature: a Defence of Literary Humanism* (OUP, 2013), and *Horace and Housman* (Palgrave Macmillan, 2013).

Allan Hazlett is Reader in Philosophy at the University of Edinburgh, having previously taught at Texas Tech and Fordham University. He received his PhD from Brown University and his main interests are in epistemology, ethics,

metaphysics, and aesthetics. He is part of a research project entitled *Intellectual virtue and the good life: Ethical and epistemic values*.

Ferenc Huoranszki is Professor of Philosophy at the Department of Philosophy of the Central European University. His interests include metaphysics and the philosophy of action, particularly the questions of free will, causation, modality, and eighteenth-century metaphysics and ethics. He is author of *Freedom of the Will: A Conditional Analysis* (Routledge, 2011).

Mark Johnston is the Walter Cerf Professor of Philosophy at Princeton University. He is the author of *Saving God* (Princeton University Press, 2009) and *Surviving Death* (Princeton University Press, 2010). Two volumes of his collected papers, *Human Beings* and *The Obscure Object of Hallucination*, will soon be forthcoming with Princeton Press.

James Ladyman is Professor of Philosophy at the University of Bristol. His work has primarily been in the philosophy of science, where the topics he has written on include scientific realism, constructive empiricism, structural realism, and the relationship between special sciences and physics. His books include *Understanding Philosophy of Science* (Routledge, 2002), and *Every Thing Must Go: Metaphysics Naturalized* (OUP, 2007) with D. Ross, D. Spurrett, and J. Collier.

Mary Leng is Senior Lecturer in Philosophy at the University of York. She has published on many topics in the philosophy of mathematics—including fictionalism, structuralism, creation and discovery in mathematics, Platonism and Anti-Platonism, and “algebraic” approaches to mathematical explanation. With M. Potter and A. Paseau she coedited *Mathematical Knowledge* (OUP, 2007). Her *Mathematics and Reality* (OUP, 2010) is a defence of mathematical fictionalism.

Jonathan Lowe was Professor of Philosophy at the University of Durham. He published over 200 articles on metaphysics, the philosophy of mind and action, the philosophy of logic and language, and early modern philosophy. His recent books include *A Survey of Metaphysics* (OUP, 2002), *Locke* (Routledge, 2005), *The Four-Category Ontology* (OUP, 2006), *Personal Agency* (OUP, 2008), and *More Kinds of Being* (Wiley-Blackwell, 2009).

Barry Loewer is Professor and Director of the Rutgers Center for Philosophy and the Sciences. He has published in philosophy of mind, metaphysics, epistemology, philosophy of logic, and philosophy of language. He is particularly known for his work on mental causation and the interpretation of quantum mechanics. He is currently writing a book on Laws, Chances, Causation and Conditionals, and coprincipal investigator (with David Albert) of a large-scale project on the Philosophy of Cosmology.

Andres Moles is Assistant Professor in both the Political Science and the Philosophy Departments at Central European University. He read Philosophy at the National University of Mexico, and received an MA in Philosophy and Social Theory (2003) and a PhD in Politics (2007) both at the University of Warwick. His research mainly concerns liberal and democratic thought, and issues concerning social and distributive justice.

Barry C. Smith is Professor of Philosophy and Director of the Institute of Philosophy in the School of Advanced Study, University of London. Most of his writing has focused on issues in the philosophy of mind and language. He coedited *The Oxford Handbook of Philosophy of Language* (OUP, 2006) with Ernest Lepore. He has been a Visiting Professor at the University of California and the Ecole Normale Supérieure.

Paul Snowdon is Emeritus Professor of Philosophy at University College, London. He was previously a tutorial fellow of Exeter College, Oxford. His research and publications relate to three main areas: the problem of personal identity (on which he is completing a book manuscript), the philosophy of perception, and the mind-body problem. He is a distinguished defender of disjunctivism in the philosophy of perception.

Daniel Stoljar is Professor of Philosophy in the Research School of Social Sciences (RSSH) at the Australian National University. He is author of two books, *Ignorance and Imagination: the Epistemic Origin of the Problem of Consciousness* (OUP, 2006) and *Physicalism* (Routledge, 2010). He received his doctorate from MIT and has been a visiting professor at Harvard.

Attila Tanyi holds a PhD (Central European University, Hungary) and is currently a lecturer in ethics and applied ethics in the Department of Philosophy at the University of Liverpool. He specializes in moral and political philosophy but his work stretches over disciplinary boundaries, and he regularly collaborates with philosophers whose specializations are very different from his own, as well as with non-philosophers with an interest in philosophical problems. He has a particular interest in experimental philosophy, and recently led the DFG (German Science Foundation) funded research group *Consequentialism and Its Demands*.

Part I

History, Methods, and Problems

1

A Different World

Barry Dainton

For anyone who has spent time immersed in, or who is a product of, the Anglophone-analytic philosophical tradition that has held sway in recent decades, since the war of 1939–45 say, the philosophical world in which Russell and Moore found themselves in the 1890s would very likely seem a very unfamiliar place. The overall tenor of recent analytical philosophy has been profoundly naturalistic and physicalistic. On this view, the fundamentals of the world are the elementary particles and fields posited by physics, which are all, without exception, entirely nonmental in nature. The dominant and most dynamic movement in mid-to-late nineteenth-century British philosophy was neo-Hegelian idealism. On this view, the world in its entirety is mental through and through. The intellectual world Russell and Moore rebelled against—the philosophical tradition to whose downfall they contributed significantly—could scarcely be more different from the one their rebellion would help to create.

The leading figures in the movement that came to be labeled “British Idealism” included Thomas Green (1836–82), Bernard Bosanquet (1848–1923), Harold Joachim (1868–1938), Francis Bradley (1846–1924), and John McTaggart (1866–1925). Although these men are often depicted as fusty reactionaries, this view of them is quite wrong: in their own day they were widely credited with revitalizing British philosophy by opening it up to important but hitherto ignored developments in the (post-Kantian) Continental tradition; the idealists viewed themselves—and were viewed by others—as something not far short of revolutionaries in their own right.

McTaggart taught in Cambridge, where Moore and Russell were both instructed—and initially inspired—by him.¹ Bradley studied and had a non-teaching fellowship at Oxford, and in the 1890s he was generally acknowledged to be the leading philosopher of his generation.² McTaggart told Moore that he believed Bradley to be the greatest living philosopher, and recalled that whenever Bradley came in “he felt as if a Platonic Idea had entered the room” (Levy 1981, p. 109). Philosophy being what it is, there were significant divergences of doctrine between the leading idealists. Green believed that God, in the form of a timeless consciousness, was immanent in us all, and that “the

unfolding of the eternal consciousness is the increasing manifestation of God in the world." McTaggart was a pluralist, and argued that reality, at bottom, is composed of interrelated immaterial selves. Doctrinal differences aside, the idealists agreed that Kant had definitively demonstrated that empiricism and related philosophies were utterly hopeless, but they also agreed with Hegel that central elements of Kant's own metaphysical system were problematic. Particularly problematic was the Kantian distinction between the phenomenal world of appearances, and the "noumenal" (unknowable) reality-as-it-is-in-itself. Hegel argued that the latter should be dispensed with, and his British followers tended to agree. Since all that remains is the *phenomenal world*, the world-as-experienced, it seems clear that reality itself is must be experiential in nature. Having established this conclusion the foundations of idealism are secure, and all that remains is to determine the precise character of this wholly mental reality, and here there are plenty of options. For our purposes, a brief overview of some of the main elements of Bradley's metaphysics will suffice. As the acknowledged leader of the Idealist movement, Bradley was the single most important figure during the period in question, and elements of his position were directly targeted by the rebels—by Russell in particular.

Bradley was a monistic "absolute" (or "objective") idealist who upheld the doctrine that reality consists of a single unified thing—the Absolute—and that this one thing is wholly experiential in nature. His metaphysic is also holistic: all the parts of the Absolute are mutually interdependent; they are such that the character of each impinges, even if only slightly, on all the rest. He also held that the Absolute was beyond our comprehension: any attempt fully to characterize it conceptually inevitably falls short, to some degree. The best we can hope to achieve are approximations. For Bradley both truth and reality always come in degrees. Naturally, he believed his own metaphysic possessed a greater degree of truth than rival systems.

Bradley arrived at this view via a combination of reason and experience.

In his metaphysics he aimed to start with as few preconceptions as possible, and arrive at a conception of reality that was maximally satisfying to the intellect.³ His major metaphysical work *Appearance and Reality* (1893) falls into two parts. The first (shorter) part is destructive: Bradley argues here that our ordinary ways of thinking about the world (both the naive and the more philosophically sophisticated) turn out to conceal contradictions when subjected to close scrutiny: "the world, as so understood, contradicts itself; and is therefore appearance, and not reality" (1893, p. 11). In the longer second part he expounds his mentalistic account of the general nature of reality. His targets in the initial destructive phase of operations include the self, space, time, motion, things, activity, and the doctrine of primary and secondary qualities. The Kantian doctrine of unknowable things-in-themselves he rejects as an absurdity: "The Unknowable must, of course, be prepared either to deserve

its name, or not. But if it actually were not knowable, we could not know that such a thing even existed . . . And this seems inconsistent" (ibid., p. 129). As for the doctrine that reality might contain components that are nonmental, Bradley found this wanting:

. . . I can myself conceive of nothing else than the experienced. Anything, in no sense felt or perceived, becomes to me quite unmeaning . . . I cannot try to think of it without realizing either that I am not thinking at all, or that I am thinking of it against my will as being experienced . . . The fact that falls elsewhere seems, in my mind, to be a mere word and a failure, or else an attempt at self-contradiction. (ibid., p. 128)

There are obvious similarities here to an famous antimaterialist argument of Berkeley's. The most original and important parts of the destructive phase of Bradley's argument are his attacks on predication and relations, and it is the elimination of the latter that opens the way to his monism.

Consider a simple thing, such as a lump of sugar. We have here an object (the lump) and its various properties (its size, shape, whiteness, etc.). The object and its properties obviously form a single whole—the propertied object—that possesses a genuine unity. But how, precisely, are we to make sense of this most familiar of situations? How do the properties and object manage to combine into a genuine whole? One option is to construe the object itself as something entirely distinct from its properties, a "bare particular," as such things are sometimes known. But these are obscure and dubious entities—it is not obvious that something entirely lacking in properties could exist—and we are still faced with the problem of how the object, construed thus, is related to its various properties. Another option is to hold that the object is nothing more than a bundle of properties. This has its merits, but it raises another issue: what is it that connects the properties to one another when they constitute an object? Once again we are confronted with the task of understanding *relations*, this time between properties themselves. The importance of relations extends well beyond the metaphysics of individual objects. The entire universe, as usually conceived, consists of a vast multiplicity of objects, and these objects are all related to one another, in multiple ways. Some are attracting others (e.g. via electrical or gravitational forces), some are impacting on others (e.g. in collisions), and more generally *all* the (nonabstract) objects in our universe are related to one another spatially or temporally—if they were not, they could not be said to exist in the same universe at all.

In chapter III of *Appearance and Reality* Bradley unleashes a dense barrage of arguments, the target of which is none other than relations in general, the glue that (we normally assume) holds the world together. His aim: to demonstrate that it is simply incoherent to think that the world could be composed

of objects or properties (“qualities” in his terminology) standing in relations to one another. There are, of course, different ways of conceiving of relations. Here, and in subsequent writings, Bradley argued that *every* way of conceiving of relations was deeply problematic. One of his arguments—often referred to simply as *Bradley’s Regress*—has acquired particular renown. One way of construing relations, in some respects a natural one, is as entities of a distinctive kind, which can exist independently of any particular objects they happen to relate. If relations are truly distinct from their relata, then if relation *R* holds between objects *O*₁ and *O*₂, say, we need a (metaphysical) account as to how and why this is so, and the only obvious way forward is to introduce further relations *R*_A and *R*_B, which connect *R* to *O*₁ and *O*₂. Since the same question now arises with respect to the connecting of *R*_A and *R*_B to *their* relata, we are embarked on an infinite explanatory regress, and Bradley—somewhat controversially—took this to mean that relations could not be real. Now, we already know that, for Bradley, reality is experiential in nature. If all relations are unreal, then reality cannot consist of a plurality of distinct but related things, hence Bradley’s conclusion: “. . . the Absolute is one system, and . . . its contents are nothing but sentient experience. It will hence be a single and all-inclusive experience, which embraces every partial diversity in concord. For it cannot be less than appearance, and hence no feeling or thought, of any kind, can fall outside its limits” (1893, p. 147).

Although Bradley held that anything approaching a full appreciation of the nature of the Absolute was beyond us, he also maintained that our own experience provides us with some important, even if partial, insights into its real nature. More specifically, the *unity* we find in our own experience is a guide to the nonrelational nature of the Absolute itself: “That on which my view rests is the immediate unity which comes in feeling” (1914, pp. 230–1). To appreciate something of what Bradley had in mind, reflect for a moment on the sort of experience one has when lying in a grassy field, looking up at the sky. One’s overall state of consciousness includes bodily feelings (the warmth of the sun on one’s skin, the tickling of the grass), together with the contents of one’s conscious thoughts (“It really is warm today” and so forth), the sound of birdsong, and the sight of the surrounding trees. These various experiences are very different in character, but they are all experienced together, as part of a single unified episode of consciousness.⁴ But although one’s visual and auditory experiences *are* unified—this much is undeniable—the unity seems to be primitive and unmediated: one is aware of the auditory and visual contents *together*, but one is not aware of any form of connecting or connective element that comes between and binds the contents. If there is a relation between these contents, it is not something that possesses any discernible experiential features it can call its own. According to Bradley at least, it is nonrelational in a further respect: we can recognize distinctions between the various parts or

aspects of our overall states of consciousness, but these parts are not objects that are capable of a separate or independent existence: “the unity of feelings contains no individual terms with relations between them, while without these no experience can be really relational” (1935, p. 642).

If the universe is a single entity, then some form of holism is difficult to avoid. If seemingly distinct objects (e.g. this table, that chair) are in reality merely aspects of *one* thing, then they are clearly not independent of one another in the way we commonly suppose. Bradley argued for a stronger form of holism, according to which the character of the whole impacts on the nature of the parts.⁵ One part of his case for this derives his logical doctrines. Bradley defended the view that all judgments—even categorical ones—are in fact abbreviated inferences. So to determine whether a judgment is true or false requires us to assess the truth of the premises of the relevant inference, and since these premises will themselves be condensed inferences—and so on ad infinitum—the truth of any one claim depends on indefinitely many more.⁶ But there was also an experiential (or empirical) underpinning to Bradley’s holism. For as we have just seen, in Bradley’s view the constituents of our total states of experience are not separable and distinct parts, and what goes for our experience also holds of the Absolute, which is itself a unified consciousness.

2 From Idealism to a Realistic (Platonic) Pluralism

Barry Dainton

Russell once observed that thinkers tend to fall into one of two categories. On the one hand there are those who view the universe as being akin to a bowl of jelly: a single, unified whole that is such that if any part of it is touched the whole quivers. On the other there are those who view the world as similar to a bucket of shot: a system that is composed of independent atoms, interrelated but capable of independent existence. Russell goes on to suggest that the most important shift in his own philosophy occurred when he rejected the holistic “jelly” view in favor of the atomistic “bucket of shot” metaphysic.¹ Russell also gave credit to Moore for opening his eyes to the failings of idealism in the 1890s and hence to the possibility of an atomistic realism. The “common sense” approach to philosophy for which Moore would become known is perhaps most famously manifest in one of his later papers, “Proof of an External World” (1939), where he establishes the existence of mind-independent material objects thus:

It seems to me that, so far from its being true, as Kant declares to be his opinion, that there is only one possible proof of the existence of things outside of us, namely the one which he has given, I can now give a large number of different proofs, each of which is a perfectly rigorous proof; and that at many other times I have been in a position to give many others. I can prove now, for instance, that two human hands exist. How? By holding up my two hands, and saying, as I make a certain gesture with the right hand, “Here is one hand,” and adding, as I make a certain gesture with the left, “and here is another”. . . . it is perhaps impossible to give a better or more rigorous proof of anything whatsoever. (Baldwin 1993, pp. 165–6)

For Moore, none of the complexities of Kant’s proof are needed. To prove the existence of the external world it suffices to hold up one’s hand and simply acknowledge what one is seeing. For Russell, being able to accept that there was a mind-independent reality—and that this reality is much as

it seems to be—was a hugely welcome liberation from the mental confines of idealism:

It was towards the end of 1898 that Moore and I rebelled against both Kant and Hegel. Moore led the way, but I followed closely in his footsteps. I think the first published account of the new philosophy was Moore's article in *Mind* on "The Nature of Judgement". . . . Although we were in agreement, I think that we differed as to what most interested us in our new philosophy. I think that Moore was most concerned with the rejection of idealism, while I was most interested in the rejection of monism. . . . They were connected through the doctrine as to relations, which Bradley had distilled out of the philosophy of Hegel . . . But it was not only these rather dry, logical doctrines that made me rejoice in the new philosophy. I felt it, in fact, as a great liberation, as if I had escaped from the hot-house on to a wind-swept headland. I hated the stuffiness involved in supposing that space and time were only in my mind. I liked the starry heavens even better than the moral law, and could not bear Kant's view that the one I liked best was only a subjective figment. In the first exuberance of liberation, I became a naïve realist and rejoiced in the thought that grass is really green, in spite of the adverse opinion of all philosophers from Locke onwards. I have not been able to retain this pleasing faith in its pristine vigour, but I have never again shut myself up in a subjective prison. (1959, pp. 54–62)

However, the conception of the world to which Moore was drawn in the initial phase of the rebellion against idealism—a conception enthusiastically shared by Russell at this time—was some distance from anything that could be described as mere "common sense."

In 1897–8 Moore wrote a fellowship dissertation "The Metaphysical Basis of Ethics" in which he argues that the Kantian conception of practical reason erroneously blurs the important distinction between our psychological faculty for making judgments and inferences, and that which is "true and objective." Moore now insists on making the sharpest of distinctions between what is mental or psychological—and this includes our judgments along with all other forms of mental representation—and the "objects of thought," the propositions themselves. Propositions may be possible objects of thought—we enter into cognitive relationships with them—but as Moore puts it in his paper "The Nature of Judgment"² this does not imply that they are themselves mental entities:

A proposition is composed not of words, nor yet of thoughts, but of concepts. Concepts are possible objects of thought; but that is no

definition of them. It merely states that they may come into relation with a thinker; and in order that they *may* do anything, they must already *be* something. It is indifferent to their nature whether anybody thinks them or not. (1899, p. 179)

So judgment involves a subject (or mind) taking up an attitude to a proposition, where the existence of the latter does not depend in any way on the apprehending mind. As for concepts, they are simply the constituents of propositions, and can be simple or complex in nature. Any complex proposition can be analyzed in terms of the simple (and not further analyzable) concepts that constitute it. Russell referred to the constituents of Moorean propositions as “terms,” and was careful to make explicit the very broad range of entities they could include: “A man, a moment, a number, a class, a relation, a chimaera, or anything else that can be mentioned, is sure to be a term; and to deny that such and such a thing is a term must always be false” (1903, §47). Since we have terms that refer to nonexistent things, these too must be real.

Importantly, relations are now also recognized, *contra* Bradley, as fully real ingredients of reality. If objects *a* and *b* are related to one another by virtue of (say) being a certain distance from one another, this fact obtains by virtue of *a* and *b*, but also the relationship of spatial distance that holds between them—relational states of affairs such as these cannot be reduced to the properties of *a* and *b* taken individually. Russell diagnosed the rejection of relations in idealists such as Bradley and Leibniz as stemming from a commitment to the view that all propositions are of subject-predicate form, a view Russell now firmly rejects. Moreover, and again *contra* Bradley, the relations Russell introduced into his ontology are “external” rather than “internal,” that is, they do not invariably affect the natures of the items they connect. The pair of objects *a* and *b* may be spatially related, but their being so does not influence the intrinsic character of either object; if *a* were at a different distance from *b*—or even at no distance from it, as would be the case if *b* did not exist—its nature would be precisely the same. And the same holds for *b*, and most other relations. This rejection of the doctrine of internal relations opened the door to the rejection of Bradleyean monism in favor of a (Platonistic) atomism; it was now an option to regard reality as being composed of interrelated but essentially independent things.³

Moore goes on to develop his Platonistic conception of propositions in a very distinctive direction. Propositions are not merely things we can think about (or with), it turns out that they—or at least the basic, atomic propositions—are the basic building blocks of the entire universe: “All that exists is . . . composed of concepts necessarily related to one another in specific manners” (1899, p. 181). For readily comprehensible reasons, this position has been dubbed “Platonic

Atomism" (by Hylton 1990). The notion that concepts constitute the very fabric of reality is a striking one, but Moore sees no alternative:

It seems necessary then to regard the world as formed of concepts. These are the only objects of knowledge; . . . A thing becomes intelligible first when it is analysed into its constituent concepts. The opposition of concepts to existents disappears, since an existent is seen to be nothing but a concept or complex of concepts standing in a unique relation to the concept of existence. (1899, pp. 182–3)

Of course, for an *idealist*, the claim that reality is composed of concepts is by no means unusual, since concepts are often construed as mental items. So one might be led to suppose that here Moore is still adhering to a form of idealism. But this would be a mistake; do not forget that by this time Moore does not think that concepts are mental entities: they are all, without exception, objective and mind-independent—though as concepts they are all also *graspable* by minds, and so in *this* sense at least they are not wholly mind-independent.

Moore registers a second deviation from idealism in the passage just cited. According to Bradley and the other followers of Hegel, compositional analysis was impossible or unilluminating; if we want to understand a thing better it is no use focusing on the smaller or more basic things of which it is composed, we need to discover how it is related to other things—and ultimately the entire universe. Moore and Russell now reject this. For as we have just seen, according to the tenets of Platonic Atomism the relationships a thing has to other things do not automatically and invariably influence its real nature; in contrast, light *is* shed on the nature of complex things by discovering the nature and identity of their more basic constituents. The rejection of internal relations means that analysis does not necessarily falsify or distort, as Bradley argued, and for this reason it can be reinstated as a viable method in philosophy.

Platonic Atomism also has consequences for the nature of truth. Whereas Bradley held that both truth and reality could come in different degrees, Moore—here too followed by Russell—maintained that there were no degrees of reality or truth. A proposition is either true, or it is false, and there are no intermediary alternatives; truth and falsity are now absolutes. As we shall see shortly, this doctrinal shift would have profound implications for the philosophy of mathematics that Russell would soon develop.

This point aside, it is natural (and common) to understand truth in terms of *correspondence*. After all, what distinguishes true propositions from false propositions is that the truths correspond with reality whereas the falsehoods do not—or so it can seem plausible to suppose. However, it is also natural to take truths themselves to be *propositions*. What is a truth if not a proposition

that is true? But if propositions are also the basic constituents of the world, it is difficult to see how truth can be a matter of correspondence. For a proposition to correspond with what is the case—the facts, as we might say—there must be a distinction between proposition and fact. But for the Platonic Atomists, there is no such distinction: propositions and what makes propositions true coincide. The true propositions simply *are* the facts.

One might conclude that Russell and Moore at this stage subscribed to an *identity* theory of truth: the doctrine that a proposition is true in virtue of being identical with the fact that makes it true.⁴ But this (to some) appealing path was blocked. The claim that “Socrates taught Plato” expresses a proposition (whose constituents include Socrates, the relation of teaching, and Plato), but so too does the claim that “Socrates is a talking cat” (one whose constituents include Socrates, and the properties of being a cat and talking). For Russell and Moore, both true and false propositions consist of complexes of objects and properties. If a proposition is true merely by virtue of being identical with the fact (or complex of objects and properties) that it constitutes, then “Socrates is a talking cat” will itself be true, since it is evidently identical with the fact it constitutes. Combining the Platonic Atomist conception of propositions with the identity theory of truth yields the absurd result that there are no false propositions at all. To avoid this, Russell and Moore eschewed the identity theory in favor of a “primitivist” account of truth; they held that truth is a basic, unanalyzable property that some propositions possess—and can be discerned to possess—but others lack. This doctrine may not have all the virtues of the identity theory, but it comes close: for it remains the case that in apprehending a proposition, we are apprehending nothing other than *what the proposition states to be the case*. Or as McDowell puts it:

... there is no ontological gap between the sort of thing one can mean ... or generally the sort of thing one can think, and the sort of thing that can be the case. When one thinks truly, what one thinks *is* the case. So since the world is everything that is the case (as [Wittgenstein] once wrote) there is no gap between thought, as such, and the world. Of course thought can be distanced from the world by being false, but there is no distance from the world implicit in the very idea of thought. (McDowell 1994, p. 27)

Some of the views adopted by Russell and Moore during the first phase of their rebellion against idealism can easily seem highly peculiar, but it is also easy to see *why* they found them attractive. Recalling these early days, Russell later wrote of “an intense excitement, after having supposed the sensible world unreal, to be able to believe again that there are such things as tables and chairs” (2009, p. 125). If one is taking this sort of delight in ordinary material objects, it would be unappealing to place barriers or intermediaries between

oneself and the world; what will appeal is a way of thinking that renders the world—the ordinary world—as accessible and intelligible. Russell made this clear in a letter to Frege in 1904:

I believe that in spite of all its snowfields Mont Blanc itself is a component part of what is actually asserted in the proposition [*Satz*] “Mont Blanc is more than 4000 metres high.” We do not assert the thought [*Gedanke*], for this is a private psychological matter; we assert the object of the thought, and this is, to my mind, a certain complex (an objective proposition, one might say) in which Mont Blanc is itself a component part. If we do not admit this, then we get to the conclusion that we know nothing at all about Mont Blanc itself.⁵

If the “theoretical want” driving one’s philosophy is the desire to explain how it is possible for us to have the direct access to the material world that we seem to have in our day to day dealings, then the position adopted by Moore and Russell will produce more intellectual satisfaction than most.

Taking a step back, although, historically, idealism succumbed to the attacks of the early analysts, it should be noted that many of the issues between them remain open or unclear to this day. Although many would argue that Bradley’s formal argument against relations, namely that they inevitably involve a vicious regress, simply refuses to acknowledge what a relation is, there is also a growing recognition that Bradley was correct in believing that there are deep and difficult metaphysical issues here—issues that are not dispelled simply by introducing predicates with two argument places into one’s logical system.⁶ It is also true that predicate logic is atomistic and extensional, in the sense that each sentence of the form Fa (which asserts that the object a has property F) is dubbed “atomic,” and the truth value of complex sentences such as “ $Fa \ \& \ Gb$ ” is a simple computation from its atomic components. But does this feature of Russell’s logic—even if it captures the nature of natural language, which is controversial—show that reality itself is atomic? After all, each number has a separate name, but no one thinks that numbers themselves are independent of each other, and the same applies for the names of the colors. The images of the world as a jelly or a bowl of shot are just images. There are many degrees in which things can be interdependent and yet distinct, and it seems likely that philosophical discussion has yet to reach the heart of this matter.

3 Principia Ethica

Barry Dainton

Moore's first book was *Principia Ethica*, published in 1903, and a development of ideas he had outlined in his 1897 dissertation "The Metaphysical Basis of Ethics." The doctrines expounded in *Principia* would be a major influence on ethical thinking in analytic philosophy in the decades to come.

Comparatively little of *Principia* is devoted to setting out Moore's views on what we should or should not do, or how we should lead our lives, or treat others. This is because his main aim is more general: "... I have endeavored to discover what are the fundamental principles of ethical reasoning; and the establishment of these principles, rather than any conclusions which may be attained by their use, may be regarded as my main object" (1903, p. ix). In more recent parlance, Moore's primary concern is *metaethical*: he wants to discover the true nature of ethics, and hence shed light on the proper way of conducting ethical reasoning and argument. The conclusions to which properly conducted ethical reasoning might lead are not his primary concern—not in this work, at least, though he does have something to say about them.

In the very first lines of the book the importance of clarification and analysis are emphasized:

It appears to me that in Ethics, as in all other philosophical studies, the difficulties and disagreements, of which its history is full, are mainly due to a very simple cause: namely to the attempt to answer questions, without first discovering precisely *what* question it is which you desire to answer. I do not know how far this source of error would be done away with, if philosophers would *try* to discover what question they were asking, before they set out to answer it; for the work of analysis and distinction is often very difficult: we may often fail to make the necessary discovery, even though we make a definite attempt to do so. But I am inclined to think than in many cases a resolute attempt would be sufficient to ensure success; so that, if only this attempt were made, many of the most glaring disagreements in philosophy would disappear. (*Preface*, p. vii)

As Moore construes it, the basic task in metaethics is determining the nature of good and evil—and hence what we mean when we say that something is good (or evil)—and this task is to be carried out by focusing on language, or more specifically, on the predicate “good,” and providing the correct definition or analysis of it. He does, however, make it clear that his is not a purely verbal or lexicographical exercise: his business is solely with the object or idea that the word is generally used to refer to. These preliminaries out of the way, the analysis Moore supplies is brutally brief, and so easily summarized: “If I am asked ‘What is good?’ my answer is that good is good, and that is the end of the matter. Or if I am asked ‘How is good to be defined?’ my answer is that it cannot be defined, and that is all I have to say about it” (1903, p. 6). This may seem innocent enough—trivial even—but if correct it has enormous implications. Utilitarians such as Mill and Bentham defined goodness in terms of happiness or well-being; for Darwin-influenced naturalists, goodness is to be defined in terms of evolutionary fitness, or success; for Kant it is a matter of acting in accord with certain rationally compelling principles; for Aristotelians, the good is essentially linked to human flourishing, and so on. Previous ethicists offered substantive analyses of the good *in other terms*. If Moore is correct, and “good” cannot be analyzed in other terms, then all preceding ethical doctrines are erroneous. Indeed, it is a fallacy—Moore labels it the “naturalistic fallacy”—to hold that goodness can be analyzed in naturalistic terms. Ethics for Moore is an autonomous, *sui generis*, discipline; as for ethical truths, they are wholly objective and mind-independent.

Moore goes on to argue that since it is indefinable, goodness must be a simple and primitive property, in the way that *yellow* is. As primitive, the property of goodness lacks parts: it cannot be decomposed into simpler or more basic ingredients. But whereas yellow is a natural property, one that can be perceived by the senses, goodness is not a natural property, and is not detectable by means of any of the five senses. Does this mean that we are unable to distinguish states of affairs that are good, and those that are evil? That would be another radical doctrine, but Moore does not subscribe to it: we *can* discern what is good and what is not, and we do so by means of a distinctive faculty of moral intuition, one that can detect the property of intrinsic goodness. Although the latter property is not a part of the natural (or material) world, it is by no means completely unrelated to it. Moore holds that situations that are indistinguishable in all physical respects will also be indistinguishable in moral respects: both will have the same overall intrinsic value. The intrinsic values of states of affairs are thus dependent upon—in later jargon, are *supervenient* upon—their (natural) intrinsic natures.

What leads Moore to this radically revisionary conception of ethics? The main line of argument is surprisingly simple. Suppose someone presents us with a putative analysis of “good,” holding—for example—that an act is good

if and only if it leads to an increase in overall happiness. Moore points out that in such a case we can still ask “Yes, but is it *good* always to act so as to maximize happiness?”; and this question is perfectly intelligible. We can conclude from this that the proposed analysis is inadequate. If the analysis *were* adequate, it would not still be an “open question” whether goodness and acting so as to maximize happiness are one and the same, but it *is* an open question, or so Moore (plausibly) maintains. The point generalizes. If we are told that something is good if and only if we desire it, does it not remain the case that we can still meaningfully ask “Is everything you desire *really* good?” Since analogous questions can be asked of every substantive putative definition of “good,” we can see that no analysis can succeed, and it follows—or so Moore argues—that goodness must be a simple and unanalyzable property.

In later chapters of *Principia* Moore turns to the application of his metaethical position. The question “How should we act, in real life situations?” turns out to have a simple answer: we should act so as to maximize the amount of goodness in the world. In practical terms this is not as helpful as it might first seem, because as Moore acknowledges, it is often difficult to know which course of action, in a given context, *will* lead to the greatest amount of goodness coming into existence. That said, as we have seen, Moore does hold that we have the capacity to detect intrinsic goodness. Where is it to be found? Moore suggests that the answer is so blatantly obvious that it risks appearing platitudinous:

By far the most valuable things which we know or can imagine, are certain states of consciousness, which may be roughly described as the pleasures of human intercourse and the enjoyment of beautiful objects. No one, probably, who has asked himself the question, has ever doubted that personal affection and the appreciation of what is beautiful in Art of Nature, are good in themselves; nor, if we consider strictly what things are worth having *purely for their own sakes*, does it appear probable that any one will think that anything else has *nearly* so great a value as the things which are included under these two heads. (1903, pp. 188–9)

The claim that beauty and personal affection are of greater intrinsic importance than anything else from an ethical standpoint is not as uncontroversial as Moore seemed to think—the elimination of poverty or pain did not rate very highly in his way of thinking.¹ But if Moore’s particular conception of where goodness lies were somewhat idiosyncratic and were to have little impact on subsequent debates, it is otherwise with his doctrines of the unanalysability of “good” and the naturalistic fallacy. His views on these matters would dominate discussions for years to come. It was not long before others, for example, Ross and Prichard were defending their own versions of (what came to be

known) as ethical intuitionism, and although the influence of intuitionism would wane, the next 75 years or so of moral philosophy were dominated by his claim that goodness could not be defined. Most of those who accepted Moore's negative argument while rejecting his intuitionism moved to some form of "noncognitivism," as the discussion of logical positivism below makes clear. Indeed, the naturalism versus anti-naturalism, and cognitivism versus noncognitivism debates still stand at the center of analytical ethics.

There is one respect, however, in which Moore's normative views have turned out to be prophetic. In §50 of *Principia* he discusses, and rejects, Sidgwick's claim that no one would suppose beauty in the natural world could have any value in the absence of any conscious human subjects who contemplate it. Moore responds with a thought experiment:

Let us imagine one world, exceedingly beautiful. Imagine it as beautiful as you can; put into it whatever on this earth you most admire—mountains, rivers, the sea; trees, and sunsets, stars and moon. Imagine all these combined in the most exquisite proportions, so that no one thing jars against another, but each contributes to increase the beauty of the whole. And then imagine the ugliest world you can possibly conceive. Imagine it simply one heap of filth, containing everything that is most disgusting to us supposing them quite apart from any possible contemplation by human being; still, is it irrational to hold that it is better than the beautiful world should exist than the one which is ugly? Would it not be well, in any case, to do what we could to produce it rather than the other? Certainly I cannot help thinking that it would. (1903, pp. 83–4)

The doctrine that the cosmos is enriched by the existence of beautiful things, even if no sentient being will ever perceive and appreciate the things in question is one that has been defended in more recent times by environmentalists.² Irrespective of what one might make of this, Moore's Platonism evidently extended well beyond a commitment to objectively real concepts and propositions.

4 Principles and Paradox

Barry Dainton

When he was 11 years old Russell was introduced to Euclidean geometry by his older brother. He made rapid progress in the subject, and later wrote “I had not imagined that there was anything so delicious in the world . . . [It was] as dazzling as first love” (2009, p. 24). His interest in mathematics continued, and he went on to take the mathematical Tripos at Cambridge, temporarily giving up philosophy so as to better prepare for his examinations. By that point Russell had been interested in the foundations of mathematics for some years. The teachers who introduced him to calculus as a teenager—the calculus then as now lay at the heart of applied mathematics—were ignorant of the recent work of important German mathematicians, such as Dedekind, Weierstrass, and Cantor, who had managed to put calculus on firm foundations, resolving issues that had tormented mathematicians since the days of Newton and Leibniz. Russell’s tutors were entirely ignorant of these recent developments:

Those who taught me the infinitesimal Calculus did not know the valid proofs of its fundamental theorems and tried to persuade me to accept the official sophistries as an act of faith. I realized that the calculus works in practice, but I was at a loss to understand why it should do so. . . . Although filled with adolescent misery, I was kept going in these years by the desire for knowledge and for intellectual achievement. . . . I hoped sooner or later to arrive at a perfected mathematics which should leave no room for doubts, and bit by bit to extend the sphere of certainty from mathematics to other sciences. (1959, pp. 27–8)

The mathematics Russell was taught as an undergraduate at Cambridge was of little assistance to him in this ambitious undertaking—he was scathing about its poor quality. But as the century drew to a close, Russell had started to familiarize himself with the work of Dedekind, Weierstrass, Cauchy, and Cantor, and he was also impressed by Whitehead’s *Universal Algebra*, which argued that mathematics was not a “science of quantity” (as was commonly supposed) but was a matter of purely formal deductive reasoning. As we

saw above, Bradley, McTaggart, and the other neo-Hegelians maintained that truth is a matter of degree, and no propositions (that are accessible to humans) are absolutely true or false. They also held that the propositions of logic and mathematics are no different from others in this regard, and pointed to various contradictions and antinomies—regarding infinity and infinitesimals, for example—as confirming evidence for this. Although Russell himself subscribed to this view during his idealist apprenticeship, as his understanding of Dedekind, Cantor et al. grew, it became increasingly evident to him that many of the alleged contradictions and puzzles had in fact been resolved by the recent advances in mathematics; Cantor’s development of the first coherent mathematical treatment of infinity was particularly significant in this respect. As Russell came to understand and accept the new mathematical findings, the hold of idealism on him gradually weakened. Also, and importantly, it was the first sign that purely technical advances in the formal sciences could have profound philosophical consequences. Russell’s enthusiasm for these developments is obvious:

From him [Zeno] to our own day, the finest intellects of each generation in turn attacked the problems, but achieved, broadly speaking, nothing. In our own time, however, three men—Weierstrass, Dedekind, and Cantor—have not merely advanced [these] problems, but have completely solved them. The solutions, for those acquainted with mathematics, are so clear as to leave no longer the slightest doubt or difficulty. This achievement is probably the greatest of which our age has to boast; and I know of no age (except perhaps the golden age of Greece) which has a more convincing proof to offer of the transcendent genius of its great men. (1918a, p. 82)

So far as the direction of his own work was concerned, the critical event occurred in 1900 when Russell attended a conference in Paris, and became acquainted with Peano’s work in logic, and its relevance to the philosophy of mathematics.¹ The goal of Peano and his Italian followers in their multi-volume “*Formulario Matematico*” (*Formulation of Mathematics* 1895–1908) was to demonstrate that all the main mathematical results of the day could be rigorously derived using the new logic they were devising for this purpose. The latter is essentially the logic based on quantifiers, variables, and propositional functions that we use today, and constitutes a considerable advance on the Aristotle-inspired syllogistic logic that it quickly replaced.² (At this point in time both Peano and Russell were ignorant of the fact that Frege had also elaborated a logic based on quantifiers and propositional functions in his *Begriffsschrift* of 1879—see §4 for more on Frege.) For Russell, the possibility that mathematics could be grounded in logic had immense appeal. For one thing, if the task could be completed it would render mathematics

epistemically secure, thus satisfying his quest for absolute certainty, and putting the legitimacy of the various mathematical resolutions of the paradoxes associated with the continuum beyond all question. It would assist in his struggle against the Idealists in a second important way, by undermining a key element of the then-influential Kantian view of mathematics. Kant held that mathematical truth is essentially connected to *intuition*, that arithmetic is grounded in the temporal features of our experience, and geometry in the spatial. Since diagrams do play an important part in some of Euclid's proofs, the latter claim at least is not without plausibility. But this view was anathema to those who—like Peano and Russell—believed the truths of mathematics to be objective, and utterly independent of the mental.

By 1900 Russell had already done a good deal of work on a book to be called *The Principles of Mathematics*, but he was not happy with it. Newly inspired by Peano's work, he immediately set about completely rewriting it. The new version was based on the new logic, which Russell quickly augmented with a treatment of relations. In an extended explosion of activity—"one of the most astonishing bursts of intense philosophical creativity in the history of the subject" (Monk 1996, p. 132)—he managed to complete a first draft of the revised *Principles* between October and December. The primary aim of the book is to show that mathematics *is* logic, that "all mathematics is deduction by logical principles from logical principles" (Russell 1903, p. 5). Since it had already been shown (in outline at least) that the rest of mathematics could be deduced from arithmetic, the key task in carrying out the "logician" project was showing that arithmetic could be reduced to logic.³ Peano had made considerable progress on this front, and shown that all of arithmetic could be logically derived from a small number of axioms.⁴ But although Peano's axioms were very basic, they include some nonlogical concepts: *zero*, *number*, *successor*. Russell set about the task of showing that these concepts could be reduced to logical ones.

In simple terms, Russell's theory of numbers runs thus. The sequence of numbers 0, 1, 2, 3, 4 . . . is identified with a *succession of classes of equinumerous classes*. So "2" is the class of all couples, "3" the class of all trios, and so forth. Of course, for this approach to be viable we need a way of identifying equinumerous classes that does not rely on the concept *number*. Russell called classes with the same number of members "similar," and defined the latter in non-numerical terms: two classes are similar if their members can all be paired off in a one-to-one fashion. The concept of a "one-to-one correspondence" is itself defined by Russell in purely logical (nonnumerical) terms: we can say that a relation *R* is one-to-one if and only if it meets the condition that if *x* and *x** bear *R* to *y*, then *x* and *x** are identical, and if *x* bears *R* to *y* and *y**, then *y* and *y** are identical. What we now need is a way of spelling out, again in purely logical terms, which classes of classes correspond with our familiar numbers

one, two, three, etc., and Russell goes on to provide precisely this. A class $N = 2$, for example, if and only if there exists an x that is a member of N , a y that is a member of N , x and y are not identical, and for any z , z is a member of N if and only if z is identical with x or with y . Three, four, and larger numbers are specified along the same lines. As for zero and one, we say that a class N has zero members if and only if there is no x that is a member of N , and it has one member if and only if there is an x that is a member of N , and for any y , y is a member of N if and only if it is identical with x .

Grounding mathematics in logic has epistemological benefits—provided at least that *logic* is secure, which Russell did not question—but the reduction of numbers to logic has *ontological* consequences as well. The ontology to which Russell subscribes in the *Principles* is the exuberant one of Platonic Atomism, which attributes *being* (if not existence) to everything we can talk about in ordinary language—creatures of fiction and myth included. Construing numbers as classes of classes allows us to dispense with numbers as objects in their own right, objects of a distinctive kind that exist over and above classes. Russell found this ontological economy congenial:

This definition has various advantages. It deals with all the problems that had previously arisen concerning 0 and 1 But much more important . . . is the fact that we get rid of numbers as metaphysical entities. They become, in fact, merely linguistic conveniences with no more substantiality than belongs to “etc.” or “i.e.”. Kronecker, in philosophizing about mathematics, said that “God made the integers and the mathematicians made the rest of the mathematical apparatus”. By this he meant that each integer had to have an independent being, but other kinds of number need not have. With the above definition of numbers this prerogative of the integers disappears and the primitive apparatus of the mathematicians is reduced to such purely logical terms, as *or*, *not*, *all* and *some*. (1959, p. 55)

The use of technical resources to eliminate one kind of entity in favor of another, more basic kind, is one that Russell would employ again and again in the years to come. It would become one of the distinguishing features of analytic philosophy.

Russell completed a revised draft of the *Principles* on December 31, 1900, and was well-pleased with himself: “In October I invented a new subject which turned out to be all of mathematics, for the first time treated in its essence. Since then I have written 200,000 words, and I think they are all better than any I have written before” (Monk 1996, p. 133). Russell’s exultant state of mind is understandable, and merited, but this “intellectual honeymoon” as he later called it was to prove short-lived. In the Spring of 1901 he discovered

a paradox involving classes that though it initially seemed trivial, he could not resolve. Since his reduction of mathematics to logic hinges on the notion of a class, the risk to his larger project was all too evident. Now known as “Russell’s Paradox,” when expressed in ordinary language it amounts to the following. Many sets, for example, the set of cats, are not members of themselves: this set is a *set*, and so not itself a cat, and so not a member of the set of cats, which includes nothing but cats. But some sets are members of themselves, for example, the set of all sets is itself a set, and so belongs to itself. We now have enough in play to state the paradox itself. Consider the set of all sets that are *not* members of themselves—call it *S*. Is this set *S* a member of itself? There are only two options, both of which are problematic. If it *is* a member of itself, we would have a problem, since as the set of all sets that are not members of themselves *S* has no members that *are* members of themselves. This leaves us with no option but to conclude that *S* is *not* a member of itself. But this too results in a problem: if *S* is not a member of itself it cannot help but find itself a member of the set of all sets that are not members of themselves, namely *S* itself.

That there was no easy solution to the problem soon became apparent. Russell discussed the issue with Whitehead, and they decided to collaborate on a new book, one that would advance the logicist program by providing a satisfactory response to the paradox, and extending the approach to additional areas of mathematics. As things turned out, this project took far longer to complete than they initially anticipated—nearly a decade longer—and the result was the three volumes of *Principia Mathematica*. But in the meantime Russell had to complete *The Principles of Mathematics* without having a fully satisfactory resolution to the paradox. What he did manage to do was provide a tentative indication of one possible solution, the “Theory of Types.”⁵ According to this theory, in broad outline, objects divide into various different types—for example, individuals, classes, classes of classes, etc.—where classes themselves are not individuals, and classes of classes are not classes of individuals, and so on. We now stipulate that no class is the right type of object to contain itself as a member. By so doing we successfully evade the paradox: the question of whether a class does or does not belong to itself cannot arise.⁶

The *Principles* was published in 1903, but his acute awareness of the unresolved paradox meant that Russell derived little pleasure from its completion. In the *Preface* he offers an apology for “publishing a work containing so many unsolved difficulties,” and the book ends with an admission and a recommendation: “What the complete solution of the difficulty may be, I have not succeeded in discovering; but as it affects the very foundations of reasoning, I earnestly commend the study of it to the attention of all students of logic” (1903, p. 528). Imperfect it may be, but the *Principles* remains an immensely

impressive achievement, and by showing how logic could impact on broader philosophical debates it is undeniably among the truly major works in early analytic philosophy. Here is Karl Popper's assessment:

The achievement of the book is without parallel. Russell rediscovered Frege's logic and theory of numbers, and laid the foundations of arithmetic and analysis. He gave the first clear and simple definition of real numbers on this basis (an improvement on Cantor, Dedekind and Peano). And he not only gave a theory of geometry but a new approach to mechanics . . . He had been the first to give a satisfactory theory of irrational numbers, a problem which had agitated mankind for 2,400 years. And he had given a really brilliant solution to the ancient problem of motion. (Magee 1971, pp. 143–5)

With the *Principles* out of the way, Russell, now in collaboration with Whitehead, plunged into the bigger work that would become *Principia Mathematica*. The final draft of this—a manuscript of some 4,500 pages—was completed in 1909, but the hundreds of pages filled with dense (and often novel) symbols posed proofreading and typesetting problems for the publishers. *Principia* was eventually published in three volumes, in 1910, 1912, and 1913. The costs were such that Russell and Whitehead each had to contribute £50 to its publication, which led Russell to joke that “We thus earned minus £50 each by ten years’ work.” These long years of arduous intellectual labor were draining, and Russell claimed he had been ruined for serious formal work by them. The demonstrations in *Principia* are so rigorous and detailed that, famously, the proof that $1+1=2$ only occurs midway through volume II.

In *Principia* classes are no longer taken as basic, but defined in terms of propositional functions. Since Russell's paradox can also be formulated in terms of these functions, the fundamental problem remains. In *Principia* a more sophisticated variant of the solution that was employed in the *Principles* is developed: the so-called Ramified Theory of Types. The additional complications this brought were unwelcome to Russell and Whitehead, but they could see no alternative—subsequent work (e.g. by Ramsey and Quine) suggests that simpler alternatives may in fact be available. In order for the new system to work, Russell and Whitehead were obliged to introduce basic axioms over and above those used in the *Principles*. One such is the “axiom of infinity,” which ensures that there is an infinite number of entities—thanks to the new theory of types, the existence of an infinite number of numbers could not be deduced logically, in the way it could in the *Principles*. Since these additional axioms were not obviously purely logical, reliance upon them undermined (somewhat) the claim that mathematics was being reduced to logic and logic alone.⁷

The viability of logicism as a philosophy of mathematics remains a controversial issue, as does the issue of what is of lasting significance in *Principia*. What is not in question is the significance of the work itself. As Hylton puts it:

Whitehead and Russell erected a unified and detailed treatment of considerable portions of mathematics. They produced the first (consistent) detailed reduction of the essential concepts of arithmetic, finite and infinite, cardinal and ordinal, and of real analysis, to the concepts of (the analogue) of set theory. They also showed that the reduction translates the truths of these branches of mathematics into the truths of (the analogue) of set theory. Whatever one may think of the underlying philosophy, and in spite of its technical flaws, *Principia Mathematica* is an extraordinary achievement. Its influence, not surprisingly, has been great. Philosophically, it played a crucial role in forming the views of Wittgenstein and of the Vienna Circle, and so directly or indirectly influenced much subsequent work in analytic philosophy. More narrowly, its technical influence is unparalleled: it founded a subject. (1990, p. 287)

The subject in question is mathematic logic, which was later developed and extended by Turing and others, and underlies the computer science that is at the heart of so much twenty-first-century technology.

5

Frege

Barry Dainton

The account of numbers sketched in Chapter 4 is generally known as the “Frege-Russell” theory, since Frege (1848–1925) had propounded the same theory some years earlier, a fact that Russell only discovered after he had just reinvented the theory for himself. In the *Principles* he gives Frege full credit, and added an appendix “The Logical and Arithmetical Doctrines of Frege,” which brought Frege’s results to wider notice, though not immediately—it was a long time before the real magnitude of Frege’s achievements dawned on the wider philosophical community. But recognition did come, and by the second half of the twentieth century, despite his narrow focus on logic and mathematics, Frege was accepted as being a figure of major importance—“the grandfather of analytic philosophy” as Dummett (1993, p. 26) puts it—and his views on logic, language, and meaning remain influential.

Frege studied chemistry, mathematics, and philosophy at the University of Jena, in Thuringia, Germany. He completed a doctorate (in which he tried to provide foundations for a part of geometry) in 1873, and received a lectureship in 1874—he would remain in Jena until his retirement in 1918. In 1879 Frege published his first significant work in logic, the *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens* (or Concept-Script: A Formula Language for Pure Thought Modeled on that of Arithmetic). The *Begriffsschrift* contained Frege’s first presentation of his new logic. It constituted an enormous advance over the existing logical systems—for example, that of Boole—but this fact largely went unnoticed by his contemporaries. This was in part due to his use of an unwieldy two-dimensional system of notation.

Frege had already become convinced that arithmetic could and should be grounded in pure logic—he was as hostile as Russell and Peano to the Kantian claim that arithmetic rested on intuition—and planned to use the system of the *Begriffsschrift* to carry this plan through, demonstrating that all arithmetical truths could be derived from the logical axioms. The realization of this plan came in stages. In 1884 he published *Die Grundlagen der Arithmetik* (The Foundations of Arithmetic), an informal work, in which he defended his own conception of arithmetic, and criticized—to devastating effect—the leading

rival views. His main positive insight here is that statements that ascribe numbers to objects or states of affairs, for example, “There are nine planets,” are really statements about *concepts*. The same portion of concrete reality can be thought about in many different ways: where one person sees a single army, another might see (perfectly correctly) ten divisions, or 50 regiments, or 500 companies, and so forth. Given this, the question “How many?” can only be answered when it is associated with a concept: we need to know if we are counting armies, divisions, or regiments. Or more generally, we need to know how many times a concept is instantiated. (1884, §46) Frege was thus led to a view similar to that which Russell would later advocate. To say that there are *two* Xs is to say that the X’s fall under the (second-level) concept of *being a concept under which two objects fall*. To avoid circularity he then provided a definition in purely logical (and so nonnumerical) terms of *being a concept under which two objects fall*. A concept *F* falls into the latter category provided it fulfills this condition: there exist distinct things *x* and *y*, *x* and *y* both fall under *F*, and if anything else falls under *F* that thing is identical either to *x* or to *y*.

Frege’s detailed work on the logicist project continued over the next decade—albeit with some setbacks as his views on some matters evolved—and in 1893 he completed *Grundgesetze der Arithmetik* (Basic Laws of Arithmetic), volume I. Here he used an expanded logical system to derive the natural numbers.¹ This was merely the first of three envisaged volumes—the second and third would embrace the real numbers and their properties. The second volume indeed appeared ten years later, in 1903. While this volume was at the printers, Frege received a letter from Russell (June 16, 1902), informing him of the class paradox, which was as devastating to Frege’s system as it was to Russell’s, as Frege quickly realized. He wrote to Russell:

Your discovery of the contradiction has surprised me beyond words and, I should like to say, left me thunderstruck, because it has rocked the grounds on which I meant to build arithmetic. . . . I must give some further thought to the matter. It is all the more serious as the collapse of my own law V seems to undermine not only the foundations of my arithmetic but the only possible foundations of arithmetic as such . . . Your discovery is at any rate a very remarkable one, and it may perhaps lead to a great advance in logic, undesirable as it may seem at first sight. (Gabriel, 1980, p. 132)

Russell was deeply impressed with the manner in which Frege received the unfortunate news:

As I think about acts of integrity and grace, I realize that there is nothing in my experience to compare with Frege’s dedication to truth. His entire

life's work was on the verge of completion . . . his second volume was about to be published, and upon finding that his fundamental assumption was in error, he responded with intellectual pleasure clearly submerging any feelings of personal disappointment. (Griffin 1992, p. 245)

Frege was forced to make hasty revisions to his second volume, modifying one of his axioms, and adding a new appendix explaining why he was doing so.² But the result was scarcely satisfactory, since the revised axiom invalidates the proofs of many of the theorems in the first volume. The projected third volume of the *Grundgesetze* never appeared, and though he continued to work, from 1904 until the time he retired, Frege published very little.

Frege's defense of logicism may have been a failure, but as George and Heck observe in their (1998) overview, there was much of lasting value in his work on logic:

In 1879, with extreme clarity, rigour and technical brilliance, he first presented his conception of rational justification. In effect, it constitutes perhaps the greatest single contribution to logic ever made and it was, in any event, the most important advance since Aristotle. For the first time, a deep analysis was possible of deductive inferences involving sentences containing multiply embedded expressions of generality. Furthermore, he presented a logical system within which such arguments could be perspicuously represented: this was the most significant development in our understanding of axiomatic systems since Euclid.

Although Frege's writings are largely technical, one of his nontechnical works has proved particularly influential in the philosophy of language: his 1892 paper "Über Sinn und Bedeutung" (On Sense and Reference). For anyone interested in arriving at a clear and simple account of how language relates to reality, a purely *extensional* account of meaning has much to recommend it: we simply say that the meaning of a name such as "Mark Twain" is the object it refers to, whereas the meaning of a predicate (or in Frege's terminology, a *concept expression*) such as "... is a writer" is the collection of objects that possess this property. However, Frege realized that this simple view is problematic, in three respects.

A sentence such as "Odysseus was set ashore at Ithaca" appears to be a perfectly meaningful. But Odysseus does not exist—at least not outside the realm of myth. If we identify the meaning of "Odysseus" with the object it refers to we face a problem: how can the whole sentence have a meaning, given that one of its principle components is without *any* meaning? Does not the meaning of a whole sentence depend on that of its constituents?

There is a second difficulty. Someone who knows little or nothing about Mark Twain other than that he was the author of *Adventures of Huckleberry Finn* would clearly learn something on being informed (and coming to believe) that Mark Twain and Samuel Clemens are one and the same person. Statements of this sort that assert identities can be, and often are, informative. It is difficult to see how they could be if the meaning of a proper name is exhausted by its reference. For if the latter is the case, how could “Twain = Clemens” differ in meaning from “Twain = Twain” or “Clemens = Clemens”? Evidently, they could not.

A third, and no doubt related problem, concerns the intersubstitutability of names in ascriptions of beliefs to individual subjects. Let us suppose a subject S is among those who know little of Twain beyond his authorship of *Huckleberry Finn*. The sentence “S believes that Twain wrote *Huckleberry Finn*” is true, we can suppose, whereas “S believes that Clemens wrote *Huckleberry Finn*” is false—S has no such belief. But once again, if the meaning of a proper name simply is the object to which it refers, the fact that “Twain” and “Clemens” refer to the same object makes it difficult to see how S could fail to possess the second belief given that he possess the first. It seems very plausible to think that expressions with the same meaning could be substituted for one another without changing the truth values of the sentences in which they occur. But in the case of sentences in which beliefs are attributed, this principle seems to be false: such substitutions do *not* always preserve truth.

Frege argued that the solution to all three difficulties lies in acknowledging that the meaning of a name is *not* exhausted by its reference. There is a further ingredient, which Frege called *sense*. The sense of the name is the *way* in which the reference of a name is presented or conceived. As it happens, “the Morning Star” and “the Evening Star” refer to the same object, the planet Venus; but the names are associated with differing modes of presentation of the planet: in the first instance as *the celestial body which rises in the morning*, in the second as *the celestial body which rises in the evening*. Frege suggested that all singular referring expressions—that is, expressions that purport to refer to a single object—have both sense and reference. Any name (in this broad sense) “*expresses its sense, stands for or designates its reference*. By means of a sign we express its sense and designate its reference.” Since different kinds of objects can be presented (or conceived) in different ways—for example, numbers and planets—the precise form sense takes will vary from case to case, but in all cases it is the sense of an expression that determines its cognitive significance for subjects who use and understand sentences in which it occurs.

The introduction of this new ingredient provided Frege with the means to solve the three puzzles. An identity such as “Twain is Clemens” can be informative because the senses of “Twain” and “Clemens” are different, even though they have the same reference. It is thus possible for “Twain is Clemens”

and “Twain is Twain” to have different cognitive significance for a subject: it is sense that determines cognitive significance, and the sense of “Twain” differs from that of “Clemens.” In the case of belief-attributions Frege held that expressions in sentences such as “S believes that the Twain is the author of *Huckleberry Finn*” the name—in this instance “Twain”—does not designate the object to which it refers, as it would in other (so-called) transparent contexts, but its ordinary *sense*. In the light of this it is easy to see how S could believe that Twain but not Clemens wrote *Finn*. Moreover, since the terms have different references, we no longer have a counterexample to the doctrine that the substitution of coreferring terms does not affect the truth-values of the sentences in which the substitutions take place.

What of sentences involving expressions that designate fictional or non-existent things, along the lines of “Odysseus was set ashore at Ithaca”? The distinction between sense and reference helps here too, since Frege held expressions could have sense even in the absence of a reference:

The words “the celestial body most distant from the Earth” has a sense, but it is very doubtful if they also have a reference. The expression “the least rapidly convergent series” has a sense; but it is known to have no reference, since for every convergent series, another convergent, but less rapidly convergent, series can be found. In grasping a sense, one is certainly not assured of a reference. (Moore 1993, p. 25)

It is not only names that have a sense and a reference for Frege, so too do complete sentences. Somewhat idiosyncratically, he held that the reference of a sentence is its truth value. Since Frege held that there are only two truth values, every sentence refers to either the True or the False. The sense of a sentence is the *thought* it expresses. Frege was insistent that (a) the sense of a whole sentence is determined by the senses of its constituents, and (b) that the reference of a sentence is determined by the references of its constituents. Frege took (b) to entail that sentences containing a nonreferring name do not determine a reference, and so have no truth value. Hence sentences such as “Odysseus was set ashore at Ithaca” are meaningful—since all their constituents have senses—but they are neither true nor false.

In one of his last important works, the 1918/19 paper *Der Gedanke* (The Thought), Frege makes it clear that Thoughts (senses of sentences) are objectively real entities, as real as tables or chairs. They are quite independent of language and mind. An experience (such as a visual experiencing of a green field) cannot exist without an experiencing subject, and it is essentially possessed by that subject—you cannot have my experience of green, and I cannot have yours. Frege’s thoughts do not need bearers or subjects: it would have been true that $2+2=4$ even if no one had ever existed to think it. Nor are

they confined to individual subjects: you and I can both think the thought that $2+2=4$. Thoughts exist in a “third realm” that is distinct from the inner world of consciousness, and the outer world of material objects.

When a thought is apprehended it at first only brings about changes in the inner world of the apprehender, yet it remains untouched in its true essence, since the changes it undergoes involve only inessential properties. There is lacking here something we observe throughout the order of nature: reciprocal action. Thoughts are by no means unreal, but their reality is of quite a different kind from that of things. And their effect is brought about by an act of the thinker without which they would be ineffective, at least as far as we can see them. And yet the thinker does not create them but must take them as they are. They can be true without being apprehended by a thinker and are not wholly unreal even then, at least if they could be apprehended and by this means be brought into operation. (Frege 1956, p. 311)

In these respects at least, Frege’s position is indistinguishable from that of Russell.

6 On Denoting, Acquaintance, and Construction

Barry Dainton

In his own quest for a satisfactory solution to the paradox that threatened the entire logicist project Russell found it necessary to change his views on reference and denoting. The resulting short but dense paper “On Denoting” (OD), which was published in 1905, proved to be his most influential and discussed contribution to the philosophy of language. In it he engages with the problems that had led Frege to introduce his distinction between sense and reference. Russell’s solution is not only very different, it also has deep and far-reaching implications for the nature and conduct of philosophy itself.

The main doctrine advanced in OD is a comparatively simple one, at least at first view. It concerns the correct logical treatment of descriptive phrases. Compare the sentences “Socrates has hair” with “The president of the USA has hair.” On the face of it, both sentences contain a subject term that serves to pick out an object (“Socrates,” “The president of the USA”), and a predicate term (“has hair”) that attributes a property to that object. Given this similarity of structure, in translating these sentences into logical symbols, the most obvious course of action would be to symbolize them both in the same sort of way. In predicate logic, basic subject-predicate sentences are taken to be combinations of names and propositional functions. Accordingly, if we let “*s*” stand for Socrates, and “*p*” for the President of the United States, and “*Hx*” designates the propositional function “. . . has hair,” then our two sentences would be symbolized thus: “*Hs*” and “*Hp*.” In this case, each of the names refers to something that exists (or existed). The situation is less straightforward when the referring expressions *do not* succeed in picking out an actual entity. Consider Russell’s own OD example:

- (a) The present king of France is bald.

The “present king of France” is a definite description—that is, it purports to refer to *one* thing—but it does not in fact refer to *anything*: there has not been

a king of France for some considerable time. So is sentence (a) true, false, or meaningless?

Since the sentence in question appears to be perfectly meaningful, avoiding the problem by declaring it to be devoid of sense is not an appealing or plausible option. The doctrines and resources of the *Principles* were such that Russell could accommodate sentences of this kind without any difficulty. Since he takes the view that the referring expressions in a sentence have to succeed in referring to something if a sentence is to be meaningful, then the definite description must have a reference. But since he also believed that there are nonexistent objects there is no problem: we need simply suppose that the present king of France is among the (many) unfortunates that possess being but not existence. In the *Principles* Russell took denoting phrases such as “the present king of France” to refer to “denoting concepts,” the meaningfulness of which depends on the existence, but not the instantiation of the relevant denoting concepts.

As his work on *Principia* got underway, Russell found himself obliged to abandon the doctrine that referring expressions in meaningful sentences invariably refer to something. After all, “the class of all classes which are not members of themselves” is a definite description, and he knew he must avoid being committed to holding that *this* expression refers to something. The solution he hit upon was to hold that the logical form of sentences containing definite descriptions—the form that most closely corresponds to the proposition they express—is very different from what their surface structure (in ordinary language) suggests. In OD Russell argues that rather than expressing a simple “referring expression + predicate” proposition, the real claim being made by (a) is this:

- (b) There is at least one thing which is the present king of France, and there is at most one thing which is the present king of France, and that thing is bald.

So according to the proposed analysis, sentences such as (a) are really a conjunction of three interrelated claims, that is, they are of the form $P = [P1 \ \& \ P2 \ \& \ P3]$. In compounded conjunctions such as P , if any of the conjuncts—that is, $P1$, $P2$, or $P3$ —is false, then the whole proposition P is also false, irrespective of the truth or falsity of the other conjuncts. In the case of sentence (b), the first conjunct (a) is simply and straightforwardly false, since there is no present king of France. If Russell is right, and (b) is an accurate analysis of “The present king of France is bald,” then the latter must also be false.

Russell uses this analysis to solve other puzzles. Consider the following claim:

- (c) Holmes is a famous detective.

Holmes does not exist, he is a fictional character, so how can this sentence be meaningful, let alone true? Whereas the Russell of the *Principles* would have invoked a nonexistent object, a Holmes inhabiting the shadowy universe of subsistents, he now sees that there is a more economical alternative. He suggests that although “Holmes” looks to be an ordinary proper name, in reality it is a definite description, a *disguised* description, something along the lines of “The fictional detective who lived at Baker Street.” We can now apply the theory of descriptions to the latter, and so analyze (c) in the same way as (a).

This approach naturally extends to the puzzles that led Frege to distinguish sense and reference. If the meaning of a proper name is the object to which it refers, then “Mark Twain is Mark Twain” and “Mark Twain is Samuel Langhorne Clemens” would have precisely the same meaning, given that Twain is in fact Clemens. But a perfectly rational person could believe the first but *not* believe the second, simply by virtue of being ignorant of any biographical facts regarding Twain. As we have seen, Frege solved this problem by introducing an additional semantic ingredient, *sense*, and holding that the objects of thought and belief are composed of senses. Russell can solve the problem more economically. We simply extend the analysis of the names of fictional and mythological characters to the names such as “Twain” and “Clemens”: we take these to be disguised descriptions. Since someone could easily associate different descriptions with each name—for example, “the author of *Huckleberry Finn*” with Twain but *not* with Clemens—we have a simple explanation as to why it is possible to believe one sentence but not the other. In OD Russell provided us with this example. George IV wished to know whether Walter Scott was the author of the novel *Waverley*. Presumably, George was not in the least curious as to whether Walter Scott and Walter Scott are one and the same person. But since Scott *did* write *Waverley*, if the meaning of a definite description is simply the object it refers to, then the king’s puzzlement would itself be puzzling: there would be no difference in (say) believing that Scott is the author of *Waverley* and believing that Scott is Scott. The problem is solved if we apply Russell’s analysis. What George IV was wondering was whether the one and only author of *Waverley* is one and the same as person as Walter Scott.

If the OD treatment of descriptions is along the right lines, then in these instances at least, ordinary language is misleading. In fact, it is misleading to a greater extent than might be obvious. Let us return to our original sentence relating to the French king, and focus on the way Russell proposes that we translate this into the language of the predicate calculus. If we let “ Kx ” mean “ x is the present king of France” and “ Bx ” mean “ x is bald” —the triad of conjoined claims in (b) is symbolized thus:

$$(d) \exists x[(Kx \ \& \ \forall y(Ky \rightarrow y=x)) \ \& \ Bx]$$

If we translate (d) back into (something approaching) ordinary language, we have something along these lines:

- (e) There exists an x which is the present king of France, and for all objects y , if y is the present king of France then y is identical with x , and x is bald.

Our original sentence “The present king of France is bald” does not *look* as though it is really a conjunction of three propositions. Nor is it readily apparent that part of what it is asserting is a *conditional* “if . . . then . . .” claim. The discrepancy between the apparent and (allegedly) real structure of sentences such as (a)—or any sentence containing referring expressions to which we apply Russell’s Theory of Descriptions—is very significant.

The approach Russell adopts in OD brings significant philosophical rewards: serious semantic puzzles are resolved and ontological economies are achieved.¹ But these gains are not cost-free, for we are obliged to accept that the structure of ordinary language sentences is an unreliable guide to the propositions they actually express. But this price may well be one many are prepared to pay, if the philosophical gains are sufficiently great. This proved to be the case. Russell’s “linguistic turn,” the bringing of the structure and analysis of ordinary language to the center of the philosophical stage, was to prove highly influential: it would not be long before analytic philosophy and the analysis of language become almost synonymous.

In one respect the linguistic turn amounts to a clear change in Russell’s earlier position. In the *Principles* the bearers of truth values are mind-independent propositions, not sentences of ordinary language, a fact that led Russell to declare that “meaning, in the sense in which words have meaning, is irrelevant to logic” (*Principles* §51). But if ordinary language was not the subject matter of logic, it was by no means entirely irrelevant to it, for at this time Russell believed that the grammar of ordinary language sentences was a largely *reliable* guide to the logical form of the propositions they express. In *Principles* §46 he writes:

The study of grammar, in my opinion, is capable of throwing far more light on philosophical questions than is commonly supposed by philosophers. Although a grammatical distinction cannot be uncritically assumed to correspond to a genuine philosophical difference, yet the one is *primâ facie* evidence of the other, and may often be most usefully employed as a source of discovery On the whole, grammar seems to me to bring us much nearer to a correct logic than the current opinions of philosophers; and in what follows, grammar, though not our master, will yet be taken as our guide.

Ordinary language grammar may have been his guide in the *Principles*, but it had lost this role by the time of OD.

In another respect, however, Russell's stance in OD (and later writings) was unchanged: he continued to adhere to the doctrine that if we understand a proposition, we must be immediately acquainted with the constituents of the proposition. However, the manner in which Russell applies this doctrine starts to evolve. In OD he argues that we are not directly acquainted with either the elementary constituents of matter posited by physics, or other people's minds. As a consequence of this, names that seem to refer to such things should be interpreted as descriptive (denoting) phrases that do not really contain the entities referred to—the particles or minds—but “contain instead the constituents expressed by the several words of the denoting phrase,” where we *are* acquainted with these. At this stage Russell was willing to accept that we are acquainted with ordinary medium-sized material objects in our surroundings—tables, chairs, other human beings, and suchlike.

But not for long. In his paper “Knowledge by Acquaintance and by Description” (1910–11) he continued to maintain that the thoughts we think are nonmental,² but he continues to stress that we can only think about things with which we are directly acquainted:

Every proposition which we can understand must be composed wholly of constituents with which we are acquainted . . . The chief reason for supposing the principle true is that it seems scarcely possible to believe that we can make a judgment or entertain a supposition without knowing what it is that we are judging or supposing about. (p. 117)

However, Russell now feels that he must further narrow the range of entities with which we are acquainted. Ordinary material objects join other people's minds and the posits of theoretical physics; the only things we are acquainted with, says Russell, are our own selves (probably) and sense-data, examples of the latter being particular sounds and regions of color as they feature in our sensory experience. It is only expressions that refer to these elemental items that are *genuine* names.³ As a corollary, he is now obliged to extend the scope of the Theory of Descriptions to nearly all ordinary language names: “Common words, even proper names, are usually really descriptions. That is to say, the thought in the mind of a person using a proper name correctly can generally only be expressed explicitly if we replace the name by a description” (1910–11, p. 114). And of course the descriptions that replace the name will, in the final analysis, contain only expressions that refer to sense-data or to oneself.

For all that Russell's position may seem extreme, it is not difficult to discern at least one of the considerations that was pushing Russell in this direction.

It is natural to think that we *know* what we are thinking or saying, that is, we are fully aware—and fully certain—of the precise content of our ordinary thoughts and utterances. Descartes argued, and Russell seemingly concurred, that we cannot be certain that the external world is as it seems—or even that it exists—but we can be certain of what we are experiencing, and of the contents of our thoughts. If we are certain of the contents of our thoughts, and our thoughts consist in our being acquainted with the constituents of the propositions we think, then the constituents of these propositions must also be things that we can be certain exist. We cannot be certain that the material bodies in our environments exist, but we can—or so Russell supposed—be certain of the existence of the sense-data we apprehend.

The position to which Russell has been led is undeniably a radical and counterintuitive one, but it was also to prove philosophically fruitful. With the foundations of mathematics secured by the *Principia*, in 1909–10 Russell started to think seriously about the physical world, and what a philosophically adequate account of it would look like. Typically, his thinking progressed quite rapidly. In the 1912 *The Problems of Philosophy* he argues that the things we are directly aware of are not the material bodies in our environments that we usually take ourselves to be aware of, but rather sense-data. The latter are situated in our private sensory spaces, and should be construed as causing our sensations. The physical objects themselves exist in a single public all-embracing “physical space”; we cannot perceive this physical space directly, but we have reason to believe—or so Russell argues—that its structure corresponds, broadly, with the spatial structures and relations we apprehend in our private sensory space. He was soon dissatisfied with this position, and in his 1914 paper “The Relation of Sense-data to Physics”⁴ he adopted a very different stance to the material world.

Taking external material objects to be the external mind-independent *causes* of our sense-data is in many respects a very natural stance—assuming one accepts that sense-data play an ineliminable role in perception—but Russell came to view it as epistemologically problematic. Any knowledge we have of material objects rests entirely on *inferences* we make from our sense-data to the (posited) external material things. What legitimates or underwrites these inferences? Since material bodies themselves are unperceivable we can never *observe* them causing sense-data, and to claim that they do is supported by no evidence of an empirical kind. Russell concluded that on this approach physics “ceases to be empirical or based upon experiment and observation alone . . . [and it] is to be avoided as much as possible.” He advocated adopting a very different approach: we do away with the unperceivable posits “by defining the objects of physics as functions of sense-data.” In essence, what Russell is proposing here is that we extend his earlier treatment of numbers to physical objects. Just as we can translate statements which

refer to numbers into statements about classes or propositional functions, we can—he now proposes—translate statements that refer to material objects into statements that refer to nothing more than sense-data. Since the latter are entirely unproblematic from an epistemological point of view, or so Russell thinks, the new translations are to be preferred, from a philosophical point of view, to the originals. He elevated this manoeuvre into an all-embracing methodological principle: “The supreme maxim in scientific philosophizing is this: *Wherever possible, logical constructions are to be substituted for inferred entities*” (1914, p. 115). He elaborated thus:

Given a set of propositions nominally dealing with the supposed inferred entities, we observe the properties which are required of the supposed entities in order to make these propositions true. By dint of a little logical ingenuity, we then construct some logical function of less hypothetical entities which has the requisite properties. This constructed function we substitute for the supposed inferred entities, and thereby obtain a new and less doubtful interpretation of the body of propositions in question. (ibid., p. 116)

Intriguingly, Russell argued that we inhabit a universe containing six spatial dimensions, consisting at it does (on his view) of the three dimensions of objective physical space, and the three dimensions of our private experiential spaces. Talk of material objects is to be replaced by talk of actual and potential sense-data (“sensibilia”); a material thing is identified with sets of sensibilia—roughly, the collection of appearances of it as perceived from different locations surrounding it: “the *matter* of a given thing is the limit of its appearances as their distance from the thing diminishes.” The structure of physical space is itself a logical construction grounded in patterns of resemblances among the contents of the private spaces. Russell himself did not develop an account along these lines in any great detail, but it would not be long before others, inspired by his “supreme maxim” took up the challenge. The two most distinguished works in this tradition, Carnap’s *Der Logische Aufbau der Welt* [*The Logical Structure of the World*] (1928/1969) and Nelson Goodman’s *Structure of Appearance* (1951, but based on work carried out in the 1930s), are among the most admired in the entire analytic canon.

Taking a step back, although Russell had reasons for moving in the direction that he did, we are now a long way indeed from where he and Moore started at the end of the nineteenth century. When Russell rejected idealism he took joy in regaining the ordinary world: “In the first exuberance of liberation, I became a naïve realist and rejoiced in the thought that grass is really green . . . I have not been able to retain this pleasing faith in its pristine vigour, but I have never again shut myself up in a subjective prison” (1959). There is surely

a degree of understatement here. Not only has Russell failed to retain his earlier faith in its pristine vigor, he is now embarked on a path that is taking him progressively further away from the world he had so recently refound, so much so that he is perilously close to confining himself within a purely subjective realm of the sort that he sought to escape.⁵

7

Wittgenstein and
the *Tractatus**Barry Dainton*

In the summer of 1911 Ludwig Wittgenstein, a young Austrian with an interest in the foundations of mathematics, paid a visit to Frege in Jena to talk about his work. At the time Wittgenstein was studying aeronautics in Manchester, but had read both Frege's *Grundgesetze* and Russell's *Principles*, and had made an attempt of his own to solve Russell's paradox. Wittgenstein later reported that Frege "wiped the floor with him" in argument, and advised him that if he wanted to pursue issues in the foundations of mathematics seriously he should go to study under Russell in Cambridge. In October Wittgenstein arrived in Cambridge, without prior warning, and presented himself to Russell.

The latter remembered Wittgenstein's arrival thus:

. . . an unknown German appeared, speaking very little English but refusing to speak German. He turned out to be a man who had learned engineering at Charlottenburg, but during his course had acquired, by himself, a passion for the philosophy of mathematics & has now come to Cambridge on purpose to hear me. (Monk 1990, pp. 38–9)

It was not long before Wittgenstein was making more of an impression: "My German friend threatens to be an infliction, he came back with me after my lecture & argued till dinner-time—obstinate and perverse, but I think not stupid" (Monk 1990, p. 39). Less than a fortnight later Russell's exasperation was increasing: "My German engineer very argumentative & tiresome. He wouldn't admit that it was certain there was not a rhinoceros in the room . . . [He] came back and argued all the time I was dressing" (ibid.). Three weeks later, with the Christmas vacation approaching, Wittgenstein was still unsure as to whether to continue working in philosophy, or return to aeronautics. As Russell relates it in his *Autobiography*:

. . . he came up to me and said: "Do you think I am an absolute idiot?" I said "Why do you want to know?" He replied: "Because if I am I shall

become an aeronaut, but if I am not I shall become a philosopher.” I said to him: “My dear fellow, I don’t know whether you are an absolute idiot or not, but if you will write me an essay during the vacation upon any philosophical topic that interests you, I will read it and tell you.” He did so, and brought it to me at the beginning of the next term. As soon as I read the first sentence, I became persuaded that he was a man of genius, and assured him that he should on no account become an aeronaut. (2009, p. 313)

Needless to say, Wittgenstein remained in Cambridge, and over the next two years or so he and Russell worked closely with one another. Their relationship was a tempestuous one—with Wittgenstein often being fiercely critical of Russell’s thinking, and vice versa. The pupil proved to be a fast learner, and within a year there were occasions when Russell felt that the future of logic was in Wittgenstein’s hands: “I love him and feel he will solve all the problems I am too old to solve—all kinds of problems that are raised by my work, but want a fresh mind and the vigour of youth. He is *the* young man one hopes for.” (Monk 1990, p. 41) One of the outstanding issues to which Russell’s work gave rise was, of course, the paradox of the class that both is and is not a member of itself. Wittgenstein did not much care for Russell’s theory of types and wanted to find a better solution. Another issue was the nature of logic itself. There is a great deal in the *Principia* that concerns the way in which mathematics can be derived from basic logical truths, but to a large extent Russell and Whitehead simply assume that logic is a secure foundation, one it is safe to build upon. But arguably, we will only be in a position to judge whether this assumption was justified when we have a clear and convincing account of the nature of logic. The nature of logical truth is another issue with which Wittgenstein concerned himself.

In October 1913, tired of what he saw as the superficial intellectual life in Cambridge, Wittgenstein took up residence in Norway, in a small village north of Bergen. The isolation proved fruitful, and his work on logic progressed well: he later recalled “. . . it seems to me that I had given birth to new movements of thought within me.”¹ Wittgenstein returned to Norway after a short Christmas break, and Moore visited him there, to report back to Russell on how his work was going. When war broke out, Wittgenstein quickly joined the Austrian army, as a volunteer, in August 1914, and remained in the military until the end of the war. After postings in artillery regiments, his request to see active service was granted, and he participated in the Kerensky Offensive (the last major Russian assault of the War in Galicia), where he was decorated for bravery; in 1918 he was transferred to the southern front in Italy, where he was captured in November of that year, detained in a POW camp at Monte Cassino, and was not released until August 1919. Throughout much of this

time Wittgenstein had managed to continue with his work on the philosophy of logic. Frege for one was astonished, as he revealed in a postcard he sent to him:

I admire your capacity for change: in the Krakow fortress, on the Weichsel with searchlights, with the field cannons, with the howitzers and now with the Hussars. And yet you still find time for scientific work! It does seem that you are more successful at that than I am.²

Wittgenstein *was* successful: by August 1918 he had managed to complete his first book *Logisch-philosophische Abhandlung* (Tractatus Logico-Philosophicus), which after some tribulation—Wittgenstein was an unknown author at the time—was published in 1921 (and in English translation shortly afterward). It would be the only book on philosophy that Wittgenstein would publish in his lifetime.³

The preface to the *Tractatus* is brief, but remarkable:

This book will perhaps only be understood by those who have themselves already thought the thoughts which are expressed in it—or similar thoughts. It is therefore not a text-book. Its object would be attained if it gave pleasure to one person who read and understood it.

The book deals with the problems of philosophy and shows, as I believe, that the method of formulating these problems rests on the misunderstanding of the logic of our language. Its whole meaning could be summed up somewhat as follows: What can be said at all can be said clearly; and whereof one cannot speak thereof one must be silent.

The book will, therefore, draw a limit to thinking, or rather—not to thinking, but to the expression of thoughts; for, in order to draw a limit to thinking we should have to be able to think both sides of this limit (we should therefore have to be able to think what cannot be thought).

The limit can, therefore, only be drawn in language and what lies on the other side of the limit will be simply nonsense.

How far my efforts agree with those of other philosophers I will not decide. Indeed what I have here written makes no claim to novelty in points of detail; and therefore I give no sources, because it is indifferent to me whether what I have thought has already been thought before me by another.

I will only mention that to the great works of Frege and the writings of my friend Bertrand Russell I owe in large measure the stimulation of my thoughts.

If this work has a value it consists in two things. First that in it thoughts are expressed, and this value will be the greater the better the thoughts

are expressed. The more the nail has been hit on the head.—Here I am conscious that I have fallen far short of the possible. Simply because my powers are insufficient to cope with the task.—May others come and do it better.

On the other hand the *truth* of the thoughts communicated here seems to me unassailable and definitive. I am, therefore, of the opinion that the problems have in essentials been finally solved. And if I am not mistaken in this, then the value of this work secondly consists in the fact that it shows how little has been done when these problems have been solved.

One thing at least is already very clear: we are dealing here with a work of quite unusual ambition. As Wittgenstein announces in the second paragraph, his concern in the book is with the problems of philosophy—not some of the problems, *all* of them—and he states in the final paragraph that he is confident that he has succeeded in solving these problems, at least in the essentials. Wittgenstein acted accordingly: having solved the problems of philosophy he decided to give up the subject. Even before his release from prison camp he had decided that he would embark on a new career as a teacher—he started training in Vienna in 1919, and took up a post in a primary school in Trattenbach in 1920.

Quite what Wittgenstein *is* proposing in the *Tractatus* is not always easy to discern. Although the book is quite a short one—a typical edition runs to less than a hundred pages which also include a substantial introduction by Russell—and nearly a century later there remains a good deal of controversy among commentators. The *form* of the book is as distinctive as its ambition: it consists of a sequence of numbered propositions, 1, 1.1, 1.12 . . . 2, 2.1, . . . 3, 3.1, 3.12, 3.121 . . . 7, where 1.1 is a comment or elaboration on 1 (which, in effect is the title of a chapter), 1.12 is a comment on 1.1, and so forth. The final proposition 7 stands alone, with no subsequent commentary. Unfortunately, the elucidatory propositions are sometimes as oracular and aphoristic as those they are supposed to elucidate. Nor is the numbering system as transparent at it initially seems: the important (perhaps key) claim “My fundamental idea is that the ‘logical constants’ do not represent; that there can be no representative of the *logic* of facts” is given the lowly number (4.0312). Wittgenstein’s seven major propositions are:

1. The world is all that is the case.
2. What is the case—a fact—is the existence of states of affairs.
3. A logical picture of a fact is a thought.
4. A thought is a proposition with a sense.
5. A proposition is a truth-function of elementary propositions.

6. The general form of a truth-function is $[p, \xi, N(\xi)]$. This is the general form of a proposition.
7. Whereof one cannot speak, thereof one must be silent.

The *Tractatus* starts off by focusing on the nature of the world (1 and 2), moves on to consider how the world is represented through thought and language (3–6), and concludes with a claim about what we must *not* talk about. The first “chapter” is so brief it can usefully be quoted in its entirety, along with the start of the second:

- 1.0 The world is all that is the case.
- 1.01 The world is the totality of facts, not of things.
- 1.11 The world is determined by the facts, and by these being all the facts.
- 1.12 For the totality of facts determines both what is the case, and also whatever is not the case.
- 1.13 The facts in logical space are the world.
- 1.2 The world divides into facts.
- 1.21 Each item can be the case or not the case while everything else remains the same.
- 2.0 What is the case—a fact—is the existence of states of affairs.
- 2.01 A state of affairs is a combination of objects (things).
- 2.011 It is essential to a thing that it can be a constituent of a state of affairs.
- 2.1012 In logic nothing is accidental: if a thing *can* occur in a state of affairs, the possibility of the state of affairs must be written into the thing itself.

Although the Tractarian metaphysics is often (and understandably) described as atomistic, it might be more accurate to say that it is *molecular*. According to Wittgenstein, reality is wholly composed of simple indivisible objects, but these objects are invariably found in combination, and these combinations are the *facts* of 1.01. The simple objects must exist, Wittgenstein goes on to argue, or else our claims about the world could not be definitely true or false; some of our claims about the world *are* definitely true (or false), therefore the simples must exist. A state of affairs is a possible combination of objects; not all possible combinations are realized in reality, so not all states of affairs are facts. The world is the totality of facts, not things, because one would not know what the world is actually like if one were simply to list the totality of objects that it contains, one also has to know how the objects are configured. States of affairs are logically independent of one another—whether or not a given state of affairs obtains has no bearing on whether any other states of affairs obtain—and in

this respect states of affairs *are* atomic in character, even though they are composite (or molecular) in nature.

Later in the (2)s Wittgenstein moves on to provide a general account of the nature of representation: his famous “picture theory” of the proposition is expounded here.

- 2.1 We picture facts to ourselves.
- 2.11 A picture presents a situation in logical space, the existence and non-existence of states of affairs.
- 2.141 A picture is a fact.
- 2.15 The fact that the elements of a picture are related to one another in a determinate way represents that things are related to one another in the same way.
 - Let us call this connexion of its elements the structure of the picture, and let us call the possibility of this structure the pictorial form of the picture.
- 2.161 There must be something identical in a picture and what it depicts, to enable the one to be a picture of the other at all.
- 2.17 What a picture must have in common with reality, in order to be able to depict it—correctly or incorrectly—in the way that it does, is its pictorial form.
- 2.172 A picture cannot, however, depict its pictorial form: it displays it.
- 3.0 The logical picture of the facts is the thought.
- 3.2 In propositions thoughts can be so expressed that to the objects of the thoughts correspond the elements of the propositional sign.
- 3.201 These elements I call “simple signs” and the proposition “completely analysed.”

So Wittgenstein is claiming that representation of worldly states of affairs is achieved via a special kind of fact: pictures. There is a special kind of correspondence—an isomorphism—between pictures and what they represent, and this extends to *thoughts*, when they are pictures of facts. What should we make of all this?

At first sight the picture theory can easily seem absurd. An ordinary sentence such as “The First World War lasted for four years” is meaningful and true, but it does not seem to contain anything that is similar in structure to the state of affairs it describes—there is no discernible trace, for example, of all the millions of people engaged in warfare in the name “First World War.” It is even less obvious that there is the “something identical” in both picture and pictured that Wittgenstein claims must exist. But Wittgenstein’s picture theory, in the form outlined above, is intended in the first instance to apply to the *elementary* propositions that describe states of affairs in a maximally

perspicuous manner. Wittgenstein believed that sentences formulated in ordinary language are (typically) in good logical order, but they do not wear their true logical form on their sleeves: "Language disguises thought. So much so that from the outward form of the clothing it is impossible to infer the form of the thought beneath it" (4.002). Ordinary language sentences can—if they really say anything about the world—be analyzed down into combinations of elementary sentences. Given the way Wittgenstein conceives of the latter, the claim that there is an isomorphism between these and the situations they refer to is *not* so absurd. A state of affairs *S* consists of certain objects in a certain combination, and these objects have certain capacities for combining with other objects. An elementary sentence *P* that successfully represents *S* will consist of names which refer to the relevant objects, and the manner in which the names are configured in *P* will represent the manner in which the objects are configured in *S*. Further, the rules of the language will be such—at least if the language is optimal—that there will be certain legitimate ways in which the names in *P* can be combined with other names to form meaningful sentences, and these will correspond precisely with the possible modes of combination of the objects to which these names refer. It is these possible modes of combination that Wittgenstein calls "form." And as we can now see, it is possible for a state of affairs and a sentence to have one and the same form.

This similarity of form is something manifest to those of us who understand sentences that represent states of affairs—it can be *shown*—but it is not something that can be pictured or described. According to Wittgenstein, all that can be pictured are the combinations of objects in states of affairs, and here the identity of form in picture and pictured is already presupposed. To describe what language and reality have in common we would have to be able to step outside language, but this we cannot do:

- 4.12 Propositions can represent the whole of reality, but they cannot represent what they must have in common with reality in order to be able to represent it—the logical form.

To be able to represent the logical form, we should have to be able to put ourselves with the propositions outside logic, that is, outside the world.

There is a deep harmony between language and reality, but this harmony is inexpressible. It is something revealed in language, but not something that can be described by language. It is this indescribable harmony—this identity of form—which makes it possible for us to say what we can say.

The bulk of the *Tractatus* is devoted to setting out Wittgenstein's distinctive (and influential) views on logic. We cannot enter into the more technical aspects of these discussions here, but we can provide an indication of the

general drift of his thinking. Wittgenstein held that logic is entirely a matter of syntactical rules relating to the use of symbols. The distinction between saying and showing applies here too. The sentences Fa and Ga each picture a state of affairs (let us suppose), and that these states of affairs obtain is what these propositions *say*. But they also *show* something; that one and the same object a occurs in both Fa and Ga . For Wittgenstein the only truths of logic are tautologies, such as “Either it is raining, or it is not raining,” or “If it raining then the shops are closed; it is raining; therefore the shops are closed.” The meaning of these sentences is such that they are true irrespective of what the world is like. Tautologies have this property by virtue of the rules governing the logical connectives—“and,” “or,” “not,” and similar expressions. These rules specify what the truth-value of the whole sentence is for every possible distribution of truth-values to its constituent parts. For example, the rules for “. . . or . . .” stipulate that “ p or q ” is *true* if both p and q are true, or if p is true and q is false, or if p is false and q is true, and the whole proposition is *false* only if both p is false and q is false. This sort of information can be elegantly presented in the form of truth-tables—Wittgenstein calls them “*schemata*” (4.31)—of the sort that are now familiar from logic textbooks. The truth table for “ p implies q ” is given below:

p	q	$p \rightarrow q$
T	T	T
T	F	F
F	T	T
F	F	T

As can be seen, the rule for “ $p \rightarrow q$ ” is such that sentences of this form are false only when antecedent (p) is true and the consequent (q) is false.

Since tautologies are true irrespective of how the world is, they do not provide us with any information; they do not make any claim about which states of affairs obtain. In this respect they are senseless. The same applies to contradictions, such as “ p and not- p ,” which are false under each assignment of truth-values to their constituents. Contradictions do not depict states of affairs: they are false no matter what, so make no claim about what the world contains. Tautologies and contradictions may be senseless, but Wittgenstein holds that they are not *nonsense*, for they succeed in showing something: “The fact that the propositions of logic are tautologies *shows* the formal—logical—properties of language, of the world.” The tautologies also constitute the *analytical* propositions (6.11)—or at least, the only analytical propositions that Wittgenstein was prepared to recognize.

Frege and Russell held that logical constants were real ingredients of reality, Platonic objects or “forms” we can mentally apprehend. Wittgenstein’s

view is more austere: he sees no reason to posit these mysterious entities. The truths of logic are without exception tautologies, and *that* they are tautologies is determined simply and solely by the syntactical rules which determine how the truth of compound sentences varies depending on the truth or falsity of their constituents.⁴ The fact that tautologies are true irrespective of what the world explains, in an economical and wholly transparent manner, their *necessity*. This is the only kind of necessity Wittgenstein acknowledges.

These conceptions of language and logic lead to the *Tractatus*' distinctive conception of the proper task of philosophy. It is not the task of philosophy to discover general truths about the universe: this is the job of the natural sciences and physics in particular. It is also a mistake to suppose that philosophers should spend their time contemplating Platonic objects or forms: Wittgenstein denies that such things exist. On one influential view, the proper task of philosophy—what distinguishes it from the natural sciences—is uncovering synthetic a priori truths, or necessary truths pertaining to reality. But Wittgenstein denies that there are such truths; as we have just seen, the only necessary truths he is prepared to acknowledge are the trivial and senseless tautologies. He does not say that metaphysics is impossible, but he does consign it to the realm of the inexpressible. Despite all this, Wittgenstein does think there is something for philosophers to do:

- 4.112 The object of philosophy is the logical clarification of thoughts. Philosophy is not a theory but an activity.
A philosophical work consists essentially of elucidations.
The result of philosophy is not a number of "philosophical propositions," but to make propositions clear.
Philosophy should make clear and delimit sharply the thoughts which are otherwise, as it were, opaque and blurry.
- 6.53 The right method in philosophy would be this. To say nothing except what can be said, i.e., the propositions of natural science, i.e., something that has nothing to do with philosophy: and then, always, when someone wished to say something metaphysical, to demonstrate to him that he had given no meaning to certain signs in his propositions. This method would be unsatisfying to the other—he would not have the feeling that we were teaching him philosophy—but it would be the only strictly correct method.

Philosophy may not be able to tell us anything about the nature of the cosmos, but it nonetheless serves the useful purpose of identifying confused ways of thinking and misleading modes of expression, even if most of these confusing statements will have been produced by other philosophers.

In 1916 Wittgenstein added this to his Notebook: “My work has extended from the foundations of logic to the nature of the world.” It is possible that by this he is referring to the way his atomistic metaphysics is rooted in the requirements of logic, or he may be referring to the propositions about *non*-logical matters that he included in drafts of the *Tractatus* around that time, possibly in response to his wartime experiences:

- 6.41 The sense of the world must lie outside of the world.
- 6.42 So too it is impossible for there to be propositions of ethics.
- 6.421 Ethics is transcendental.
- 6.431 So too at death the world does not alter, but comes to an end.
- 6.44 It is not *how* things are in the world that is mystical, but *that* it is.

There is more in the same vein. Quite what Wittgenstein meant by these claims, and how they relate to the rest of the *Tractatus*, has been much debated ever since he wrote them.

The final two propositions of the *Tractatus* are as remarkable—and controversial—as any of their predecessors:

- 6.54 My propositions are elucidations in the following way: anyone who understands me eventually recognizes them as nonsensical, when he has used them—as steps—to climb up beyond them. (He must, so to speak, throw away the ladder after he has climbed up it.) He must transcend these propositions, and then he will see the world aright.
- 7 What we cannot speak about we must pass over in silence.

Whatever else it might be, the *Tractatus* is not a work of natural science, nor is it filled with tautologies. Much of it is filled with what looks to be ordinary philosophy—at least if we overlook the brief mystical excursions—albeit expressed in an uncommonly concise manner. Yet according to the account of meaning Wittgenstein expounds and defends in the book, philosophical claims such as these are nonsense and the sentences in which they are formulated do not express genuine propositions, but are at best “pseudo-propositions.” It is only in the realm of natural science that genuine fact-stating propositions are to be found. There are also the tautologies, but although these are true, necessarily so, they are senseless, even if they are not nonsense. As Wittgenstein reveals in 6.54, he was well aware of the paradoxical predicament in which he has left us: if the central claims in the main body of the *Tractatus* are true, then they are nonsense; but if they are nonsense, then obviously they cannot be true. Russell’s paradox threatened to undermine the logicist project. Wittgenstein’s paradox threatens to undermine the whole of philosophy—*itself* included.⁵

The *Preface* ends by Wittgenstein claiming that the value of his book consists in its showing how little is achieved when the problems of philosophy have been solved. It seems he was intent on demonstrating that everything of real value—ethics, aesthetics, the meaning of life and death—lies in the realm of the inexpressible, and he has accomplished precisely this by showing precisely where the limits of the sayable lie.⁶ Be that as it may, the problem remains: why should we believe the limits of sense and nonsense are where Wittgenstein claims, if the arguments that take him to this conclusion are senseless? Did Wittgenstein believe that the philosophical propositions set out in the *Tractatus* are a privileged kind of nonsense, of a sort which can lead one to the truth? That is one possibility, but before accepting it without qualification there is another claim made in the *Preface* that should not be overlooked: “the *truth* of the thoughts communicated here seems to me unassailable and definitive.” Fully reconciling this statement with the doctrine that philosophical propositions are nonsensical is not easy: surely statements that are true—definitively and unassailably so—cannot also be nonsense. Perhaps we can say no more than this. Wittgenstein clearly believes that the positive accounts of logic and language that he defends—and argues for—in the *Tractatus* are superior to all the alternatives, the theories of Frege and Russell included. His account of these matters is the finest that philosophy can achieve, and so is *as* true as any positive philosophical or metaphysical theory can be. The doctrines of the *Tractatus* may strictly speaking be nonsense, but they are not *just* nonsense: in Wittgenstein’s eyes at least, they are where reason ultimately leads, they are the ultimate destination of philosophy. It is the misfortune of philosophy—and reason itself—that these doctrines are ultimately self-undermining. This may seem regrettable, tragic even, but this is simply how it is. Or so Wittgenstein then believed.⁷

8 The Vienna Circle and its “Wissenschaftliche Weltauffassung”

Barry Dainton

In the Autumn of 1924 Wittgenstein left Trattenbach to start work in a new primary school in Ottertal. On December 25 of the same year Moritz Schlick wrote to him, asking for permission to visit him and discuss his ideas. Schlick was the leader of a discussion group in Vienna that was taking great interest in the doctrines of the *Tractatus*, which had been published in German in 1921. This association of scientists, mathematicians, and philosophers would become known as the “Vienna Circle.” The Circle itself was comparatively short-lived: it began in 1924, and most of its members started to leave Austria and Germany when the Nazis came to power in 1933; Schlick was murdered in 1936. However its influence was extensive and enduring. It gave birth to logical positivism—also known as *logical empiricism*¹—which would become the next main strand or phase of analytic philosophy: it became the dominant form in the United States, which after the Second World War would become the main center of analytic philosophy itself.²

The Circle was a somewhat loose association, whose membership changed over time. Aside from Schlick (holder of a chair in the philosophy of science at the University of Vienna whose previous holders include Ernst Mach and Ludwig Boltzmann), its other leading members were Otto Neurath and Rudolf Carnap, but the membership included many eminent figures: Friedrich Waismann, Herbert Feigl, Gustav Bergmann, Hans Hahn, Philipp Frank, Karl Menger, Richard von Mises, and Kurt Gödel.³ The Circle viewed themselves as revolutionaries, or at least, as attempting to forge a revolutionary way of doing philosophy.⁴ The late nineteenth and early twentieth centuries had seen a great flourishing in the fundamental sciences. Maxwell’s theory brought the first comprehensive understanding of electricity and magnetism. Boltzmann had revolutionized thermodynamics. Einstein had overturned Newton’s conception of space, time, and gravity—and with his General Theory of Relativity exploited non-Euclidean geometry, producing a crisis in neo-Kantian circles.⁵ Inspired by these developments, the positivists wanted to develop a way of

doing philosophy that would revolutionize the subject in the way that Einstein and others had revolutionized physics. The new philosophy would be resolutely empiricist in character. Knowledge of reality would be grounded entirely in experience. Metaphysics of the obscurantist or speculative variety would have no place in the new scheme of things. Mathematics and logic had always proved problematic for radical empiricists; it is not easy to demonstrate that *they* are grounded in experience. But the positivists were hopeful that progress on this front was now possible, thanks to the new work of Hilbert, Russell, Whitehead, and Wittgenstein. The Vienna Circle found the latter's *Tractatus* a particular inspiration: it was read aloud and discussed line by line.⁶ The Circle did not agree with Wittgenstein on all matters, but they agreed with much of the *Tractatus*, including the doctrine that logical truths are tautologies.

In a 1929 public manifesto the Circle gave a name to the new approach they were seeking to develop. The title of the manifesto was "*Wissenschaftliche Weltauffassung: Der Wiener Kreis*" (The Scientific World-conception: The Vienna Circle), and it began with these lines:

Many assert that metaphysical and theologising thought is again on the increase today, not only in life but also in science. Is this a general phenomenon or merely a change restricted to certain circles? The assertion itself is easily confirmed if one looks at the topics of university courses and at the titles of philosophic publications. But likewise the opposite spirit of enlightenment and *anti-metaphysical factual research* is growing stronger today, in that it is becoming conscious of its existence and task. In some circles the mode of thought grounded in experience and averse to speculation is stronger than ever, being strengthened precisely by the new opposition that has arisen.

In the research work of all branches of empirical science this *spirit of a scientific conception of the world* is alive.⁷

They go on to list kindred spirits, past and present, including J. S. Mill, Mach, Bolzano, Boltzmann, Brentano, Meinong, Mally, Menger, Adler, Poincaré, Duhem, Einstein, Peano, Frege, and Wittgenstein—and note that "anti-metaphysical endeavours" are particular flourishing in England, where the recent logical work of Russell has reinvigorated the great local tradition of empiricism.

As with any philosophical school there were variations in doctrine among the individual logical positivists, and their doctrines evolved in response to criticism; the views of Carnap changed at an almost Russellian rate. But for present purposes we can overlook most of these variations and evolutions, and concentrate on the core doctrines that most of the Circle subscribed to, at least initially.

We saw in the *Preface* that for Schlick the conception of philosophy Wittgenstein had set forth in the *Tractatus* was a decisive event in the history of the subject, and he embraced it wholeheartedly, as he explains in “The Turning Point in Philosophy,” the opening essay of the first issue of the positivists own journal, *Erkenntnis* in 1930: “philosophy is that activity through which the meaning of statements is revealed or determined. By means of philosophy statements are explained, by means of science they are verified.”⁸ The clarification of scientific and other meaningful discourse is the primary positive contribution the philosopher can make to human knowledge. With this focus on the meaning of statements, the “linguistic turn” initiated by Russell and Wittgenstein had evidently found new adherents. Carnap thought philosophers could play a somewhat more positive role: he held that the *construction* of artificial languages could assist in the project of putting science on firm foundations, and clarifying its real content. But Carnap too rejected the idea that it was the job of philosophy to teach us anything about the nature of reality. Metaphysics in its traditional guise was to be shunned.

The Circle’s hostility to metaphysics was grounded in part in a radical empiricism that took experience to be the only guide to knowledge, in part in a theory of meaning. They held that there are just two forms of meaningful sentences: those that express analytic statements, and those that make genuine factual claims about the world. They viewed analytic truths as being rooted in conventions governing the use of symbols. Logical truths are tautologies that tell us nothing about the world, and are true in virtue of the conventional meaning of the logical constants. Arithmetic—as Russell and Whitehead had shown—can be reduced to logic and is thus analytic, and they were confident the rest of mathematics would prove amenable to the same treatment. Nonlogical analytic truths, such as “Bachelors are unmarried” are simply stipulative definitions. They held that all analytic truths are necessary and a priori. Necessary because they could not be false—given that our conventions are what they are—and a priori because we do not need to consult the world in order to know our own conventions. The analytic truths are the *only* truths that are either necessary or a priori.

Since the positivists restricted the a priori to the analytic, there was no place in their system for *synthetic* a priori truths, that is, truths about reality that we can arrive at independent of empirical investigation. They thus ruled out one of the traditional sources of metaphysical knowledge. But there was a further important strand to their attack on metaphysics, deriving from the second type of meaningful statement they were prepared to recognize: those that make genuine factual claims about the world. The positivists placed an unusually stringent requirement on factual statements. They held that for a nonanalytic statement to be meaningful (or cognitively significant) it must be *empirically verifiable*. To understand a purportedly factual claim we must (a)

know the conditions in which it would be true, and (b) know the empirical data—the observations—which would enable us to ascertain whether it is true or false. This is the positivists' famous—notorious—"verification principle."

The idea that understanding a factual claim involves knowing what the world would have to be like if the claim were true is relatively noncontroversial, not so the doctrine that we also have know how to verify the claim. If genuinely meaningful statements of the nonanalytic variety are limited to those that are empirically verifiable then vast regions of traditional philosophical and theological discourse will turn out to be meaningless. In his 1932 paper "The Elimination of Metaphysics Through Logical Analysis of Language" Carnap supplies examples of the kinds of terms and expressions one can find in "pseudo-statements," that is, sequences of words which look like genuine statements at first glance but are in reality devoid of meaning; the list includes "principle," "God," "the Absolute," "emanation," "absolute spirit," and "the Ego." For examples of complete pseudo-statements Carnap turns to some quotations from Heidegger's *What is Metaphysics* (1929): "We assert: *the Nothing is prior to the Not and the Negation . . .* Where do we seek the Nothing? How do we find the Nothing? . . . *What about this Nothing?—The Nothing itself nothings.*"⁹ For Carnap metaphysical statements such as these are disguised gibberish, not profundities worthy of prolonged study. Filled with revolutionary zeal as they were, the members of the Circle were unperturbed by their virtual elimination of a subject; indeed since they viewed it as progressive, they welcomed it. But there is no denying that the philosophy of the scientific world-conception is as narrow as it is austere.

The verification principle might have been a useful tool in the struggle against traditional metaphysics, but it turned out to be a troublesome ally. In A. J. Ayer's *Language, Truth and Logic* (1936) the book that brought logical positivism to the attention of the Anglophone world and fame to its 26-year-old author, we are told that "We say that a sentence is factually significant to any given person, if, and only if, he knows how to verify the propositions which purport to express it" (p. 35). But what, precisely, does "verify" mean here? This is a question that both exercised and divided members of the Circle. The verification was held to consist in the first instance of reports of observations—"protocol statements" as they were sometimes called—but there was a divergence of view on the character of these. Some held that protocol sentences were confined to claims about the contents of immediate sense-experience, others (e.g. Neurath) held that they could concern physical objects. Ayer himself went on to distinguish between verifiable *in principle* and verifiable *in practice*, and between *strong* (conclusive) and *weak* (probabilistic) verifiability. In both cases Ayer favored the more liberal alternative. Statements about (say) the kind of microorganisms that are to be found on the planets orbiting distant stars seem to be perfectly meaningful, but it may well never be a practical proposition for us to go and verify them. By holding that such statements are to be regarded

as verifiable *in principle*, the liberal positivist can take them to be as contentful as they seem. Universal generalizations such as “a body tends to expand when heated” are commonplace in science—most scientific laws take this form—but they too are under threat from the verification principle. To be absolutely certain the generalization holds we would need to observe each and every occasion on which a body is heated, which is obviously nonfeasible—not least because many such instances occurred before there were *any* observers in the universe at all. So Ayer again opts for the more liberal (and plausible) position. The question we need to ask is not whether any observations would make the truth (or falsity) of a statement logically certain, but “Would any observations be relevant to the determination of its truth or falsity?” (1971, p. 38) Finding a way of formulating the verification principle that was both clear in what it entailed, but not open to logical difficulties, proved difficult—more than 30 years later Ayer himself conceded that “I’m not sure to this day that all the gaps have been mended” (p. 54).¹⁰

In *Principia Ethica* Moore claimed to be putting forward a “Prolegomenon to any future ethics that can possibly pretend to be scientific” (1903, p. ix). Moore may have held that moral truths were wholly objective, but the existence of such truths depended on a property of intrinsic goodness that could only be detected by a special faculty of intellectual intuition. Unsurprisingly, the positivists in general, and Ayer in particular, were not inclined to treat the elusive deliverances of this mysterious faculty as equivalent to ordinary observations. Ayer goes on to argue that even among those who claim to have this faculty, its deliverances are often incompatible, with differing moral philosophers arriving at differing views as to the appropriateness or otherwise of the same course of action. Ayer acknowledges that some intuitionist ethicists claim to *know* that their moral judgments are correct, but he is unmoved:

... such an assertion is of purely psychological interest, and has not the slightest tendency to prove the validity of any moral judgment. For dissenting moralists may equally well “know” that their ethical views are correct. And as far as subjective certainty goes, there will be nothing to choose between them. When such differences of opinion arise in connexion with an ordinary empirical proposition, one may attempt to resolve them by referring to, or actually carrying out, some relevant empirical test. But with regard to ethical statements, there is, on the “absolutist” or “intuitionist” theory, no relevant empirical test. We are therefore justified in saying that on this theory ethical statements are held to be unverifiable. (1971, p. 141)

If ethical judgments are empirically unverifiable, then if they are to have genuine meaning within the positivists’ framework, they will have to be analytic,

or tautologous. But are they? Ayer argues not. Suppose someone holds that an action is right (or a state of affairs good) because it is generally approved of. We can see at once that such an analysis fails, since it is not self-contradictory to assert that there are some actions that are generally approved that are wrong (or that some states of affairs that are generally judged to be good, are in fact bad). In a similar vein, suppose someone were to hold that an action is morally right only if it leads to a greater quantity of well-being in the world than all the other possible course of action. Once again, since the denial of this is not self-contradictory—one might easily and intelligibly hold that it is sometimes wrong to act in a way that maximizes well-being—the proposed analysis fails: the moral claim in question is not an analytic truth. And since the same goes, says Ayer, for all the other accounts of moral judgment that with which he is familiar, we have no option but to conclude that the ethical statements are nonanalytic. Since they are also nonempirical, this means that they are not asserting genuine propositions. Despite appearances to the contrary, moral assertions are devoid of cognitive significance, in just the same way as their metaphysical counterparts, and like the latter they cannot be true or false.

This does not mean that there cannot be a scientific inquiry into values, but this will take the form of psychological and sociological investigation into the attitudes people in a given society, at a given time, actually have, and the underlying psychological—and perhaps neurophysiological—processes that lead to this state of affairs. Or as Ayer puts it: “The propositions which describe the phenomena of moral experience, and their causes, must be assigned to the science of psychology, or sociology.” Schlick anticipates and responds to an objection:

One might wish to derive from this a supposedly profound and destructive objection to our formulation of the problem. For, one might say, “in such a case there would no ethics at all; what is called ethics would be nothing but a part of psychology!” I answer, “Why shouldn’t ethics be a part of psychology?” . . . if we decide that the fundamental question of ethics, “Why does man act morally?” can be answered only by psychology, we see in this no degradation of, nor injury to, [moral] science, but a happy simplification of the world picture. In ethics we do not seek independence, but only the truth. (1939, pp. 29–30)

The positivists were still left with a problem to solve. If, as they claimed, statements that express ethical judgments are devoid of genuine factual meaning, what distinguishes them from meaningless noises? Since they do not seem meaningless, we need a plausible account of how and why moral expressions are used as they are. Ayer succinctly summarized his stance on this issue thus: “in so far as statements of value are significant, they are ordinary ‘scientific’

statements [in psychology etc.]: . . . in so far as they are not scientific, they are not in the literal sense significant, but simply expressions of emotion which can be neither true nor false." (1971, p. 136). If you see someone stealing X's money, and say to them "You acted wrongly in stealing X's money," according to Ayer, on the factual level you have not said anything more than "you stole money from X." When you state that this action was wrong, you are not making any further factual claim about it, you are simply conveying that you disapprove of this action. And this disapproval—according to Ayer—amounts to no more than this: the action provoked or caused certain (negative) *feelings* in you. This account applies quite generally: ". . . in every case in which one would commonly be said to be making an ethical judgment, the function of the relevant ethical word is purely 'emotive'. It is used to express feelings about certain objects, but not to make any assertion about them" (ibid., p. 143).

Ayer acknowledges that there is an additional element, one also emphasized by Reichenbach and Carnap. Ethical statements do express our emotions, but they often do more: they are also deployed to arouse feelings in others, with a view to moving them to act in certain ways in response to these feelings. There is an entire spectrum of possibilities in this regard. An assertion of "It's your duty to tell the truth" might well be an expression of a feeling possessed by the speaker, but also a command, to the effect: "Do your duty!" In contrast "You ought to tell the truth" also involves a command, typically, but the tone is discernibly milder or less emphatic. More generally, Ayer suggests, the meaning of the various ethical words can be defined "in terms both of the different feelings they are ordinarily taken to express, and also the different responses which they are calculated to provoke" (ibid., p. 143).

This "nonscognitivist" or "emotivist" (as it came to be known) account of ethics has some obvious advantages. It renders the mysterious unnatural properties and the associated faculty of "intuition" posited by Moore entirely redundant. Also, it is clear that moral evaluations often do involve feelings, and moral claims are often made with a view to influencing actions, even if that is not always the sole or most important intention. In subsequent decades more sophisticated variants of it would be developed—notably by C. L. Stevenson in a succession of influential works, including "The Emotive Meaning of Ethical Terms" (1937), and *Ethics and Language* (1944)¹¹—and the positivists' approach to ethics was widely embraced in analytic philosophy. Although in due course the tide turned and emotivism fell from favor, other forms of nonscognitivism continued to flourish: Blackburn's "quasi-realism" and Gibbard's "norm-expressivism" are two important contemporary variants of this approach.¹²

Ethics aside, a further interesting commitment of the positivists was to the "Unity of Science Programme," a particular interest of Neurath's. Proponents

of the Dilthey-inspired view of the *Geisteswissenschaften* held that the physical and human sciences (including psychology and sociology) had fundamentally different methodologies, the former essentially involving hermeneutics, the latter not. The positivists were hostile to doctrines of this kind, and insisted that *all* the sciences shared—or should share—a common methodology: “The goal ahead is *unified science* . . . The endeavor is to link and harmonize the achievements of individual investigators in their various fields of science” (Manifesto). To this end Neurath planned on organizing and publishing a vast *Encyclopedia of Unified Science*, in 26 volumes, each of ten monographs. Bringing together the work of all the sciences was a vast task, but the Circle were confident that they had the basic framework in which it could be brought about:

The aim of the scientific effort is to reach the goal, unified science, by applying logical analysis to the empirical material. Since the meaning of every statement of science must be statable by reduction to the given, likewise the meaning of every concept, whatever branch of science it may belong to, must be statable by step-wise reductions to other concepts, down to the concepts of the lowest level which refer directly to the given. If such an analysis were carried through for all concepts, they would thus be ordered into a reductive system, a “constitutive system.” (“Manifesto,” Neurath 1973, p. 309)

This passage, which goes on to claim that the most basic level of the system will consist of experiential concepts, that from these are constituted the “layer” of physical objects, then other minds and the “objects of the social sciences” bears the fingerprints of Carnap in his *Aufbau* period. Shortly afterwards Carnap abandoned reductive phenomenalism, and, following Neurath’s lead, accepted that protocol sentences referred to ordinary physical objects, rather than private sensations. His aim now was to show that the language in all the sciences—both high-level and fundamental theoretical physics—could be reduced to statements about these physical objects. Although the unity-of-science project could not be deemed a success, the issues it raised relating to the reducibility (or nonreducibility) of different scientific theories remain central topics in the philosophy of science, and the later work of logical empiricists such as Hempel on the structure of scientific explanation, and Nagel on inter-theoretic reduction, became standard works in the field.¹³

The young Karl Popper attended several meetings of the Circle, but although he quickly won their respect, he was never asked to join—probably because Schlick was unimpressed by the ferocity of his hostility to Wittgenstein’s views.¹⁴ In 1934 Schlick did however publish Popper’s *Logik der Forschung*—translated into English as *The Logic of Scientific Discovery* in 1959—in the “Monographs on the Scientific World-Conception” collection

that he coedited with Frank. Although Popper had the same deep respect for the sciences as any official member of the Circle, his philosophy of science, as presented in his *Logik*, was distinctive. Whereas the Circle appealed to the verifiability criterion as a way of distinguishing sense from nonsense, Popper believed this task was futile—indeed, he thought the topic of meaning was itself philosophically unimportant. In its stead he offered a criterion of “demarcation” that would distinguish genuine science (such as Einstein’s relativity theories) from pseudo-science (such as astrology or some forms of psychoanalysis). Controversially, at least by the lights of the Circle, Popper believed Hume’s problem of induction was insoluble, and that the quest for a logic of induction was thus futile. No matter how many times we have seen the sun rise in the morning, we have no more reason to think it will rise when the next morning comes than we had to start off with. If inductive confirmation does not distinguish scientific hypotheses from the ungrounded speculations of pseudoscience, what does? For Popper *falsifiability* is the hallmark of genuine science. Putative scientific laws can never be conclusively confirmed by empirical evidence, but they can be refuted by it: if the sun does not rise tomorrow, the universal claim “the sun will always rise” has been decisively shown to be wrong. Hence for Popper the distinguishing feature of science is a willingness to submit one’s theories to empirical testing, and abandon them if the results are negative. The debate between positivist philosophers of science and Popper on the viability of induction and inductive logic would continue for decades to come.¹⁵

Of all the positivists, Carnap arguably made the largest impact on analytic philosophy taken as a whole.¹⁶ The *Aufbau* (1928) is among the enduring monuments of the heyday of logical positivism. In it we find Carnap attempting to carry through a phenomenalistic reductionism—roughly, translating all meaningful statements into statements about sense-data—with unprecedented rigor and technical precision. He takes as his starting point momentary cross-sections of entire streams of consciousness, and using a single primitive relationship of “remembered similarity” (between sense-data at different times), he seeks to construct everything else. He later adopted a conventionalist-cum-pragmatic stance on choice of base-level; hence the adoption of “physicalism” noted above. In 1934 Carnap published *The Logical Syntax of Language* (an English translation appeared in 1937), where he deployed sophisticated new metalinguistic methods inspired by Tarski—involving an infinite hierarchy of languages—to demonstrate that we can coherently and consistently talk about languages *within* language, and hence that the Wittgensteinian doctrine of “saying versus showing” can be dispensed with.¹⁷ Following Tarski’s successful development of a model-theoretic theory of truth, Carnap published several works on semantics, including his *Introduction to Semantics* in 1942. He then turned to modal logic, and his *Meaning and Necessity: A Study in Semantics*

and *Modal Logic* (1947) was an important contribution to that topic. His later work focused on induction and probability theory, resulting in works such as the *Logical Foundations of Probability* in 1950 and *The Continuum of Inductive Methods* in 1952.

Popper was not the only critic of the positivists' conception of science. In 1962 Thomas Kuhn's *The Structure of Scientific Revolutions* appeared, and would soon make waves. Kuhn argues that Popper's own account of how science should be done is not confirmed by several key phases in the history of science, where we generally find scientists doing their best to hang on to their (major) theories when confronted with potentially damaging empirical results. Kuhn also outlined an influential picture of science being conducted within systems of unquestioned assumptions or "paradigms." Since different paradigms employ different modes of reasoning, the transition between them cannot be fully rational, and scientific progress is itself not as rational as had previously been thought.

Although Kuhn's views on the nature of science are themselves sometimes presented as involving a radical break with the past, the situation is more nuanced, at least with regard to Carnap. Kuhn's book was under the aegis of the Circle's "Unified Science" series, and there are very definite similarities between Kuhn's paradigms and Carnap's doctrine of "linguistic frameworks," as expounded in the latter's "Empiricism, Semantics and Ontology" (1950):

Are there properties, classes, numbers, propositions? In order to understand more clearly the nature of these and related problems, it is above all necessary to recognize a fundamental distinction between two kinds of question concerning the existence or reality of entities. If someone wishes to speak in his language about a new kind of entity, he has to introduce a system of new ways of speaking, subject to new rules; we shall call this procedure the construction of a linguistic *framework* for the new entities in question. And now we must distinguish two kinds of question of existence: first, questions of the existence of certain entities of the new kind *within the framework*; we call then *internal* questions; and second, questions concerning the existence or reality *of the system of entities as a whole*; called *external* questions. Internal questions and possible answers to them are formulated with the help of the new forms of expression . . . An external question is of a problematic character which is in need of closer examination.

For Carnap the choice between linguistic frameworks cannot be rationally defended in the same way as internal questions.

Although the positivists may have been primarily concerned with establishing a scientific world-conception, they also had broader ambitions, some

of which go to explain the messianic quality to some of their writings. In the preface to the *Aufbau* we find this:

We do not deceive ourselves about the fact that movements in metaphysical philosophy and religion which are critical of such [a scientific] orientation have again become very influential of late. Whence comes our confidence that our call for clarity, for a science that is free from metaphysics, will be heard? It stems from the . . . belief that these opposing powers belong to the past. We feel that there is an inner kinship between the attitude on which our philosophical work is founded and the intellectual attitude which presently manifests itself in entirely different walks of life; we feel this orientation in artistic movements, especially in architecture, and in movements which strive for meaningful forms of personal and collective life, of education, and of external organization in general. We feel around us the same basic orientation, the same style of thinking and doing. It is an orientation which demands clarity everywhere, but which realizes that the fabric of life can never quite be comprehended. It makes us pay careful attention to detail and at the same time recognizes the great lines that run through the whole. It is an orientation which acknowledges the bonds that tie men together, but at the same time strives for the free development of the individual. Our work is carried by the faith that this attitude will win the future.

Carnap here gives voice to the wider social and cultural aspirations of the Vienna Circle. It goes without saying that the early 1930s was an immensely fraught period in European history, and the positivists were inevitably influenced by the turmoil in the Weimar republic and the rise of Nazism.¹⁸ They hoped that their determination to achieve clarity and transparency in philosophy and science would be part of a broader transformative movement with similar ultimate goals.¹⁹ If with the benefit of hindsight this optimism seems naïve, it should surely not be held against them.

9 Later Wittgenstein

Barry Dainton

Wittgenstein gave up his school teaching career in 1926, and in between working on the design of the Stonborough house and doing some sculpture, he gradually resumed contact with other philosophers—he spoke with various members of the Vienna Circle, and the young Cambridge philosopher Frank Ramsey, who travelled to Vienna to meet with him on several occasions.¹ In 1929 Wittgenstein himself returned to Cambridge, where the *Tractatus* was accepted as a doctorate,² and he embarked on new philosophical work, producing a sizeable manuscript in a few months. It was on the basis of this that he was awarded a research scholarship at Trinity College in 1930, and he began lecturing. When assessing this new work Russell wrote:

The theories contained in the work of Wittgenstein are novel, very original and indubitably important. Whether or not they are true, I do not know. As a logician who likes simplicity, I should wish to think that they are not, but from what I have read of them I am quite sure that he should have an opportunity to work them out, since when completed they may easily prove to constitute a whole new philosophy.³

Wittgenstein would remain in Cambridge—where he was appointed chair of philosophy in 1939—until his resignation in 1947.

Russell's assessment proved to be accurate: from that point until his death in 1951 Wittgenstein *did* go on to develop a whole new philosophy, often referred to as simply that of “the later Wittgenstein.” The new philosophy—or new *approach* to philosophy—was not developed overnight. His new thinking was initially triggered when Ramsey persuaded him that there were some irreparable flaws in the *Tractatus*.⁴ Quite rapidly this led him to think that his early views were radically misguided in some important respects. As he elaborated and exploited the new approach over the next few years Wittgenstein wrote a great deal, but published very little. The *Philosophical Investigations*—the culmination of this second phase of his career, and regarded by many as his masterpiece—appeared two years after his death, in 1953. The *Preface* to this work is noticeably less forceful and confident than its counterpart in

the *Tractatus*, concluding thus: “I should like to have produced a good book. This has not come about, but the time is past in which I could improve it.” He also tells us “I should not like my writing to spare other people the trouble of thinking. But, if possible, to stimulate someone to thoughts of his own.” In this at least the work has been an unqualified success. Although superficially much more accessible than the *Tractatus*—we still have numbered propositions but the hierarchical decimal system has gone, logical symbols feature scarcely at all—working out quite what Wittgenstein was trying to say about the many issues with which he engages in the course of the book is no easy matter. Wittgenstein himself describes the thoughts expressed in the book as “the precipitate of philosophical investigations” that have occupied him for the past 16 years, and as he observes,

They concern many subjects: the nature of meaning, of understanding, of logic, the foundations of mathematics, states of consciousness, and other things. I have written down all these thoughts as *remarks*, short paragraphs, of which there is sometimes a fairly long chain about the same subject, while I sometimes make a fairly sudden change, jumping from one topic to another. . . . The philosophical remarks in this book are, as it were, a number of sketches of landscapes which were made in the course of these long and involved journeyings.

When reading the *Investigations*, one seldom has any trouble understanding the individual paragraphs or remarks. The trouble starts when one tries to work out what one is meant to learn from them—what point or purpose they serve—for since Wittgenstein rarely spells this out, there is no alternative to thinking for oneself. Needless to say, there is considerable scope for different readings of Wittgenstein’s intentions, hence in the years following the publication of the *Investigations* there has been a healthy industry in Wittgensteinian exegesis.⁵

We cannot even outline here Wittgenstein’s contributions to the many issues with which he deals in the *Investigations* and other writings (in Part II Richard Gaskin discusses one important aspect of his thought—on rule following). We can, however, say a little about his new method, and the associated conception of philosophy itself, which would soon prove to be influential. In a lecture given in the early 1930s, Wittgenstein claimed that philosophy as he was now practicing it was not merely a stage in the continuous development of the subject, he claimed to be working in a *new form* of philosophy, deploying a new method, one that makes it possible to have skilled philosophers as well as great ones. What was this new method?

In the *Tractatus*, Wittgenstein made a number of claims about the nature of philosophy. It is an a priori discipline, one that cannot *say* anything about

the world. Philosophical problems arise from our misunderstanding the logic of our language, which occurs because certain grammatical forms—forms of words—mislead us, and it is the task of the philosopher to clarify these misleading propositions. But Wittgenstein also held that “all the propositions of our everyday language, just as they stand, are in perfect logical order” (TLP 5.5563). There is an obvious tension here, one that was exacerbated by the fact that he did not actually provide any complete analyses of ordinary language sentences. By 1930, Wittgenstein had started to have doubts of his own concerning his earlier views:

The following is a question I constantly discuss with Moore: Can only logical analysis explain what we mean by the propositions of ordinary language? Moore is inclined to think so. Are people therefore ignorant of what they mean when they say “Today the sky is clearer than yesterday”? Do we have to wait for logical analysis here? What a hellish idea! Only philosophy is supposed to explain to me what I mean by my propositions and whether I mean anything by them. (McGuinness 1967, pp. 129–30)

The new method, as it developed over the next few months and years, is broadly in agreement with the conception of philosophy elaborated in the *Tractatus*, although there are also important divergences. The proper targets of the philosopher’s attention are *philosophical problems* themselves; Wittgenstein still thinks that philosophical problems are rooted in confusions, often brought on by language misleading us in some way or other. He continues to hold that it is not the job of philosophy to describe the world by producing scientific theories—we can leave that to the scientist—and metaphysics is still to be shunned. But there is a crucial difference. Although philosophical confusions are to be eliminated by the obtaining of an improved perspective on language, this will not be achieved, as it was in the *Tractatus*, by revealing an invisible substructure underling ordinary language. The focus is on grammatical forms that are visible to anyone who looks—the job is to describe how we actually *use* the words in our ordinary everyday language. Once this is done, and if it is done well, the confusions that create philosophical difficulties will be swept away. In 1931 Wittgenstein tells us that he used to be of the view

. . . that the task of logical analysis is to discover the elementary propositions . . . Only in recent years have I broken away from that mistake . . . The truth of the matter is that we have already got everything, and we have got it actually *present*; we need not wait for anything. We make our moves in the realm of the grammar of our ordinary language, and this grammar is already there. Thus we have already got everything and need not wait for the future. (McGuinness 1967, p. 183)

Wittgenstein held that philosophical questions and answers typically stray into subtly deviant uses of words. The questions philosophers ask—and the answer they propose—are frequently nonsensical, so by getting clear on how words are actually used we can clarify the bounds of sense and nonsense; and when we do, the nonsensical nature of the original philosophical puzzle will be obvious. Of course, we all know how to use words perfectly well; we do not speak nonsense in daily life (very often). Philosophers can mostly speak as well as anyone else in their day-to-day lives. It is only when they enter their studies and start doing philosophy that the trouble starts.

This gives rise to an inevitable question. If all we are trying to do is describe how language is actually used, will we not be engaged on the same task as the linguist or grammarian? Wittgenstein denies this: the philosopher's interest in language-use is *therapeutic*. The victim of a philosophical puzzle is confused about certain specific words and their uses, and the goal is to cure this particular puzzlement, by discovering its various sources. The portions of grammar that need to be studied to bring about the cure will be highly distinctive—rather than a description of how we ordinarily use words, which is the concern of the linguist, the focus is on various subtle *misuses* of words (in philosophy).

PI 127 The work of the philosopher consists in assembling reminders for a particular purpose.

PI 132 We want to establish an order in our knowledge of the use of language: an order with a particular end in view; one out of many possible orders; not *the* order. To this end we shall constantly be giving prominence to distinctions which our ordinary forms of language easily make us overlook.

Although we have all learned the correct use of words, we are not explicitly aware of the various *differences* in the way words are used in different contexts. When we first encounter mathematicians and philosophers saying that numbers are names of objects, and that these objects can be proved to *exist*, most of us are somewhat puzzled, but we come to accept it nonetheless. Since differences between the ways we ordinarily use the names of numbers and the names of ordinary physical things are easily overlooked, we simply assume that numbers and tables exist in just the same way. And this leads us to think that mathematical knowledge and discoveries are just like empirical knowledge and discoveries, albeit with one difference: mathematicians are exploring not the ordinary world of physical things, but a strange world of abstract entities. Of course, this picture of mathematics leads to all sorts of intractable puzzles and problems that philosophers such

as Plato, Frege, and Russell have been trying to solve ever since. According to Wittgenstein, if we had paid more attention to the differences of usage here it would soon have become apparent that “exists” means something quite different when applied to numbers than what it means when applied to tables and chairs, and much needless bafflement and effort could have been avoided.

On Wittgenstein’s new approach, philosophy consists of rearranging things we already know; it is rather like rearranging familiar books in a library. Analysis in the *Tractatus*-style was more of an adventure: the search was for an ideal notation that would capture the concealed underlying form of every possible language. The new approach is a good deal less exhilarating.⁶ From now on, Wittgenstein will still be concerned with “language” and “words” for much of the time, but he will now be concerned with *these* words, as they are used in *our* language. Confusions about the nature of language will be solved by looking at how we actually talk about language, meanings, explanations of meanings, sentences, and so forth:

PI 120 Your questions refer to words; so I have to talk about words.

PI 97 Thought is surrounded by a halo—Its essence, logic, presents an order, in fact the a priori order of the world: that is, the order of *possibilities*, which must be common to both world and thought. But this order, it seems, must be *utterly simple*. It is *prior* to all experience, must run through all experience; no empirical cloudiness or uncertainty can be allowed to affect it.—It must rather be of the purest crystal. But this crystal does not appear as an abstraction; but as something concrete, indeed, as the most concrete, as it were the *hardest* thing there is. (TLP 5.5563)

We are under the illusion that what is peculiar, profound, essential in our investigation, resides in its trying to grasp the incomparable essence of language. That is, the order existing between the concept of proposition, word, proof, truth, experience and so on. This order is a super-order between—so to speak—super-concepts. Whereas, of course, if the words “language,” “experience,” “world,” have a use, it must be as humble a one as that of the words “table,” “lamp,” “door.”

In this passage—and many others in the early sections of the *Investigations*, he is attempting to correct the mistakes to which he himself succumbed in his earlier work.

Wittgenstein was aware that his new method would not satisfy those addicted to the supposed depths of metaphysics, engaged in the quest for

essences, truths about transcendent realities, and so forth. But if such philosophizing is merely peddling nonsense, we must force ourselves to remain content with the modest descriptive method—which is very difficult to employ in practice. The reward is peace of mind, of a kind.

PI 118 Where does our investigation get its importance from, since it seems only to destroy everything interesting, that is, all that is great and important? (As it were all the buildings, leaving behind only bits of stone and rubble.) What we are destroying is nothing but houses of cards and we are clearing up the ground of language on which they stand.

PI 119 The results of philosophy are the uncovering of one or another piece of plain nonsense and of bumps that the understanding has got by running its head up against the limits of language. These bumps make us see the value of the discovery.

PI 133 . . . the clarity that we are aiming at is indeed *complete* clarity. But this simply means that the philosophical problems should *completely* disappear.

The real discovery is the one that makes me capable of stopping doing philosophy when I want to.—The one that gives philosophy peace, so that it is no longer tormented by questions which bring *itself* in question. . . .

There is not *a* philosophical method, though there are indeed methods, like different therapies.

PI 309 What is your aim in philosophy? To show the fly the way out of the fly-bottle.

Several of the distinctive themes and doctrines associated with Wittgenstein's later philosophy make an appearance in the very early sections of the *Investigations*. The book opens with a quotation from Saint Augustine in which the latter outlines an account of how language is learnt. The passage from Augustine is quite rich, but includes the idea that learning a language involves learning which objects various words signify, and then using these words to express desires. Wittgenstein describes it thus:

These words . . . give us a particular picture of the essence of human language. It is this: the individual words in language name objects—sentences are combinations of such names. In this picture of language we find the roots of the following idea: Every word has a meaning. This meaning is correlated with the word. It is the object for which the word stands.

He then immediately introduces a simple *language-game*, a primitive way of using language.

You send someone shopping, with a slip of paper marked “five red apples.” The shopkeeper takes the slip, opens the drawer marked apples; looks up the word “red” in a table, and finds a colour-sample opposite. He then says out aloud the series of numbers 1–5, and with each number takes a red apple out of the drawer.

Wittgenstein comments:

It is in this and similar ways that one operates with words. “But how does he know where and how he is to look up the word ‘red’ and what he is to do with the word ‘five?’” — Well, I assume that he *acts* as I have described. Explanations come to an end somewhere—But what is the meaning of the word “five”?—No such thing was in question here, only how the word “five” is used.

Several distinctive themes are already in play. Wittgenstein wants to draw attention to the fact that words do not just *name* objects, but are interwoven with various forms of activity, for example, we use words to get people to *do* all manner of things. The example also usefully illustrates that words do not necessarily have to have meanings in the form of objects they denote. In the simple shopping game, “apple” may refer to physical objects, but not any particular object; “red” does not refer to an object at all, but a color; “five” does not refer to an object at all—the shopkeeper has learned a certain procedure, which hearing the word “five” triggers. How does the shopkeeper know how to act? How does he know what the words mean? All manner of explanations might occur to us, for example, that his understanding of the words involves a mental process of interpretation, and it is this which ultimately leads to his action. But he may just have been *trained* to respond in this way to these words—in essentially the way we train a dog—he just sees the words and *acts* without anything more going on at all. The claim that explanations “come to an end,” and often terminate in simple behavioral patterns is a recurrent theme in the later Wittgenstein’s writings.

Wittgenstein does not introduce the term *language-game* until a little later (PI 7): as well as artificially simple uses of language, such as already given, he says games like ring-a-ring-a-roses are language-games . . . also “I shall also call the whole, consisting of language and the actions into which it is woven, the ‘language-game.’” Later still, in PI 23, he says:

. . . There are *countless* different kinds of use of what we call “symbols,” “words,” “sentences” . . . the term “language-game” is meant to bring into

prominence the fact that the *speaking* of language is part of an activity, or of a form of life.

Review the multiplicity of language-games in the following examples, and in others:

- Giving orders and obeying them
- Describing the appearance of an object . . .
- Constructing an object from a description or drawing
- Reporting an event . . .
- Speculating about an event . . .
- Forming and testing a hypothesis
- Making up a story
- Play acting
- Singing catches
- Guessing riddles
- Making a joke
- Solving a problem in practical arithmetic
- Translating from one language into another
- Asking, thanking, cursing, greeting, praying . . .

It is interesting to compare the multiplicity of the tools in language and of the ways they are used, the multiplicity of kinds of words and sentence, with what logicians have said about the structure of language. (Including the author of the *Tractatus Logico-Philosophicus*.)

Again it is apparent that language-games are uses of language in a general way of life that involves people doing far more than just describing what they see, and it seems he is willing to count each different use of language (jokes, prayers, orders) as a different language-game. Wittgenstein anticipates an objection to this whole approach:

PI 65 Here we come up against the great question that lies behind all these considerations.—For someone might object against me: “You take the easy way out! You talk about all sorts of language-games, but have nowhere said what the essence of a language-game, and hence of language, is: what is common to all these activities, and what makes them into language or parts of language. . . .”

And this is true.—Instead of producing something common to all that we call language, I am saying that these phenomena have no one thing in common which makes us use the same word for all,—but that they are *related* to one another in many different ways. And it is because of this relationship, or these relationships, that we call them all “language.” I will try to explain this.

PI 66 Consider for example the proceedings that we call “games.” I mean board-games, card-games, ball-games, Olympic games, and so on. What is common to them all? Don’t say: “There *must* be something common, or they would not be called ‘games’” —but *look and see*, whether there is anything common to all. —For if you look at them you will not see something that is common to *all*, but similarities, relationships and a whole series of them at that. To repeat: don’t think, but look! —Look for example at board-games, with their multifarious relationships. Now pass to card-games; here you find many correspondences with the first group, but many common features drop out, and others appear. When we pass next to ball games, much that is common is retained, but much is lost. —Are they all “amusing”? Compare chess with noughts and crosses. Or is there always winning and losing or competition between players? Think of patience. In ball games there is winning and losing; but when a child throws his ball at the wall and catches it again, this feature has disappeared. Look at the parts played by skill and luck; and at the difference between skill in chess and skill in tennis. Think now of games like ring-a-ring-a-roses; here is the element of amusement, but how many other characteristic features have disappeared! And we can go through the many, many other groups of games in the same way; can see how similarities crop up and disappear.

And the result of this examination is: we see a complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail.

PI 67 I can think of no better expression to characterize these similarities than “family resemblances”; for the various resemblances between members of a family: build, features, color of eyes, gait, temperament, etc. etc. overlap and criss-cross in the same way. — And I shall say: “games” form a family.

. . . But if someone wished to say: “There is something common to all these constructions — namely the disjunction of all their common properties” — I should reply: Now you are only playing with words. One might as well say: “Something runs through the whole thread — namely the continuous overlapping of those fibres.”

So, there is no single feature common to all games. The concept “game” cannot be analyzed in terms of necessary and sufficient conditions: there is no property that an activity has to have in order for it to be a game, there is no property the possession of which guarantees that an activity is a game. There is no *essence* that something must have to be a game. All that unites the

various games that exist is a web of similarities. The concept “game” is what is now known as a *family-resemblance* concept. Many concepts are of this kind, as becomes clear as soon as we think about them.

The concept “language” is one such, or so Wittgenstein suggests. If there is nothing all languages have in common, languages do not have in common what the *Tractatus* says they have in common. If language has many uses, if it works in fundamentally different ways, then no single account of it will suffice. Does this mean logic does not apply? Of course not. Logic applies perfectly well to certain regions of language, but we must not expect (or demand) that it apply to every proposition. That we must not squeeze language into a single ill-fitting garment is but one of the many lessons Wittgenstein would have us learn from his *Investigations*.

The change of orientation and approach that characterized the second main phase of Wittgenstein philosophical career was not to everyone’s taste. Russell saw little or no merit in it:

There are two great men in history whom [Wittgenstein] somewhat resembles. One was Pascal, the other was Tolstoy. Pascal was a mathematician of genius, but abandoned mathematics for piety. Tolstoy sacrificed his genius as a writer to a kind of bogus humility which made him prefer peasants to educated men and *Uncle Tom’s Cabin* to all other works of fiction. Wittgenstein, who could play with metaphysical intricacies as cleverly as Pascal with hexagons or Tolstoy with emperors, threw away this talent and debased himself before common sense as Tolstoy debased himself before the peasants—in each case from an impulse of pride. I admired Wittgenstein’s *Tractatus* but not his later work, which seemed to me to involve an abnegation of his own best talent very similar to those of Pascal and Tolstoy. (1959, p. 159)

As Russell was well aware when he wrote this, in the late 1950s, this low opinion of Wittgenstein was not shared by many of the younger generation of philosophers.⁷

10 Quine

Barry Dainton

A. J. Ayer was not the only English-speaking foreigner to meet with the Vienna Circle in the early 1930s: so too did a young philosopher from Akron Ohio in the United States, Willard van Orman Quine. Having recently completed his PhD at Harvard—on aspects of *Principia Mathematica*—Quine was travelling in Europe for a year, on a scholarship. After spending six weeks in Warsaw with Tarski he moved on to Prague, to study under Carnap. There, as Quine himself put it, he soon became a “disciple” of Carnap for the next few years, and discovered “what it was like to be intellectually fired by a living teacher rather than by a dead book” adding “I had not been aware of the lack” (Creath 1990, p. 465) Not that Quine ever seemed greatly concerned with the issues that had troubled the philosophers of earlier eras. Schlick and Carnap started their philosophical careers immersed in German metaphysics and reacted against it, Russell and Moore were fighting against an idealism they had initially made their own. Quine had no sympathy whatsoever with the sorts of metaphysical systems developed by neo-Kantians and Hegelians, but as Hylton observes: “those were not battles that he felt a need to fight” (2007, p. 33).¹

Much of Quine’s early work was in logic and set theory. Of *Principia Mathematica* itself he later commented “This is the book that has meant the most to me.” But although he greatly admired the symbolic parts of the work, he was less impressed by the “explanatory patches of prose that were interspersed” in it:

I was taken with the clear, clean incisiveness of its formulas. But this was not true of its long introduction to volume 1. . . . In those pages the distinction between sign and object, or use and mention, was badly blurred. Partly in consequence, there was vague recourse to intensional properties or ideas These ill-conceived mentalistic notions paraded as the philosophical foundations for the clean-cut classes, truth functions and quantification that would have been better as a starting point in their own right. (1991, p. 265)

This preference for the clear and clean-cut, the hostility to the vague talk of “ideas”—and an insistence on distinguishing use and mention—would be characteristic of Quine’s philosophy over the coming decades. In his dissertation he set about clarifying and improving the foundations of *Principia*, eliminating the need for “intensional entities” such as propositions and properties in favor of sentences and objects; his hostility to the intensional would also remain intact throughout his long career.

Quine may have regarded himself as a disciple of Carnap but one of his earliest influential contributions was an attack on one of his master’s most cherished doctrines. The positivists tended to follow the lead of Frege and Russell in reducing mathematics to logic. Anyone who embarks on this path is left with the problem of understanding logic and logical truth. In his *Logical Syntax of Language* (1934) Carnap sought to establish that logical truth is nothing more (nor less) than linguistic convention: it amounts to a choice of syntax for a language in which our scientific theories can be formulated. In “Truth by Convention” (1936) Quine argues that this cannot be right. The elementary truths of logic generate, deductively—that is, via logical inference—an infinite number of further (more complicated) logical truths. Given this, conventionalists have an uncomfortable choice to make. They could hold that there is a different and distinct convention, one for each logical truth. Since there is an infinity of logical truths, and humans are finite beings, this is obviously implausible—it is, after all, *we* who establish the conventions. But the other alternative is just as unwelcome. If there are only a finite number of conventions, the only way to generate the infinite number of logical truths that are *not* created by conventions is by using logic itself, and so we have not succeeded in showing that logic can be reduced to convention. Or as Quine himself puts it: “In a word, the difficulty is that if logic is to proceed *mediately* from conventions, then logic is needed for inferring logic from the conventions.”²

Quine’s own philosophy could be summed up by the maxim *take science seriously*. Look around and ponder for a moment while considering this question: What is our best source of knowledge of reality? What is providing us with greater insight into the nature of things than anything else? Is there in fact a putative provider of knowledge that stands out from the crowd? For Quine it is perfectly obvious that there is: natural science, construed broadly so as to include biology, history, chemistry, cosmology in addition to fundamental physics. His naturalistic standpoint is on vivid display in this passage from “The Scope and Language of Science”:

I am a physical object sitting in a physical world. Some of the force of this physical world impinge on my surface. Light rays strike my retinas; molecules bombard my eardrums and fingertips. I strike back, emanating

concentric air waves. These waves take the form of a torrent of discourse about tables, people, molecules, light rays, retinas, air waves, prime numbers, infinite classes, joy and sorrow, good and evil. My ability to strike back in this elaborate way consists in my having assimilated a good part of the culture of my community, and perhaps modified and elaborated it a bit on my own account. . . . Now how is it that we know that our knowledge must depend solely on surface irritation and internal conditions. Only because we know in a general way what the world is like, with its light rays, molecules, men, retinas, and so on. (1957/1976, pp. 228–9)

Since the Vienna Circle was also inclined to take science seriously, the maxim “take science seriously” is by no means a uniquely individuating feature of Quine’s position. What *is* distinctive about Quine’s naturalistic philosophy is what he does with the maxim. Arguably he takes it further than anyone before—or since.

Some of the most distinctive elements of Quine’s naturalism result from his divergences from the doctrines of the logical positivists. The latter held that there is a strict distinction between analytic and synthetic truths, taking the former to be necessary and a priori, and the latter to be a posteriori and contingent. Quine came to see things differently. In what is probably his most influential single paper, “Two Dogmas of Empiricism” (1951), Quine argues that none of the accounts of analyticity that philosophers have thus far provided are acceptable, and that the concept (at least as normally construed) should play no role in serious philosophy. Potentially, a great deal hangs on the outcome of this argument. In establishing that there is no viable concept of analyticity Quine also argues, *en route*, that there is no viable concept of *synonymy* (sameness of meaning); but if there is no such thing as sameness of meaning, the concept of meaning itself is under threat—as Quine fully appreciates and intends. And that is only the beginning. For the positivists the a priori and the *necessary* truths are not only related to the analytic truths, they *depend* on them, so if analyticity goes, so too do the a priori and the necessary truths.

So the stakes are potentially high. Quine begins his attack by focusing on the accounts of analyticity provided by Kant and Frege, and argues that they fail because they either presuppose the notion of analyticity at some point in their analyses, or they rely on concepts that are just as problematic as analyticity—for example, synonymy, correct definition, necessity—in their explication of it. In the final parts of the paper Quine moves on to consider whether the concepts of analyticity and synonymy might thrive in the context of the verificationist conception of meaning. The natural thing for a verificationist to say is that two statements are synonymous if they would both be verified by the same range of experiences, and analytic if they are true *come what may*, that is,

true irrespective of what happens in the world. Quine seeks to undermine this way of construing analyticity too, deploying *en route* a holistic “network” conception of language and scientific theory that has proved to be highly influential subsequently.

Quine proposes that we view scientific theories as interconnected webs of sentences. Since he sees common sense and science as continuous with one another, our ordinary everyday beliefs are, for him, part of science—a humble part, but science nonetheless—the network model applies to them too. We are thus led to view our total set of beliefs, ranging from the most banal trivialities concerning ourselves, our pasts and our environments, to the most exotic, concerning the latest developments in science, as an interconnected system of sentences that we hold true: a “web of belief.” It is useful, Quine suggests, to think of the network as having an exterior surface and an interior. The exterior surface is where sensory experience impinges: it consists of those sentences we are most inclined to cease holding true in the light of new sensory experience. These may be loosely called “observation sentences,” for example, “there is a cat on the mat” or “the sky is overcast today.” As we move inwards from the periphery, we come to more “theoretical” sentences. Sensory experience can impact on these too, but only indirectly, sometimes very indirectly, and there are many different ways of making the necessary accommodations. Here is how Quine himself puts it:

Total science is like a field of force whose boundary conditions are experience. A conflict with experience at the periphery occasions readjustments in the interior of the field. Truth values have to be redistributed over some of our statements. Reevaluations of some statements entails reevaluation of others, because of the logical interconnections—the logical laws being in turn simply further statements of the system, certain further elements of the field. Having reevaluated one statement we must reevaluate some others, which may be statements logically connected with the first or may be the statements of logical connections themselves. But the total field is so underdetermined by its boundary conditions, experience, that there is much latitude of choice as to what statements to reevaluate in the light of any single contrary experience. (1953, pp. 42–3)

As is clear from the final claim, Quine subscribes to the doctrine of “underdetermination of theory by data.” On this view, scientific theories are not logically entailed by the empirical data which support them, and there is always a multiplicity of *different* theories that can accommodate and explain any given set of data. This is why the total field is underdetermined by what occurs at the boundary.

Quine goes on to suggest that the interior/exterior distinction is a soft one: no claims, not even the logical and mathematical statements at the heart of our belief system, are completely immune to revision in the light of new experiences:

It is misleading to think of the empirical content of an individual statement—especially if it is a statement at all remote from the experiential periphery of the field. Furthermore it becomes folly to see a boundary between synthetic statements which hold contingently on experience, and analytic statements, which hold come what may. Any statement can be held true come what may, if we make drastic enough adjustments elsewhere in the system. Even a statement very close to the periphery can be held true in the face of recalcitrant experience by pleading hallucination or by amending certain statements of the kind called logical laws. Conversely, by the same token, no statement is immune to revision. Revision even of the law of excluded middle has been proposed as a means of simplifying quantum mechanics; and what difference is there between such a shift and the shift whereby Kepler superseded Ptolemy, or Einstein Newton, or Darwin Aristotle? (1953, p. 43)

While Quine believes every statement that we accept at a given time is in principle revisable, he also says that in practice we manage our belief-system by aiming to minimize overall change. As theorizers we are economical: when, for example, experience conflicts with current belief, or new mathematical discoveries force theoretical change, we try to accommodate the conflict by making the most minimal changes to the existing system as we can. But occasionally, we are forced to revise our theories in the light of intolerable conflict with data: Kepler's theories did finally displace Ptolemy's, as Newton's replaced Kepler's. The required changes could even be such as to require us to revise our most basic laws of logic. We thus see that, for Quine, the distinction between analytic and synthetic is not an absolute one. Some sentences are more immune to revision in the light of future experience or theoretical developments, but no sentences, not even "*p* or not-*p*," are *entirely* immune to revision. There are no analytic truths in this sense. Nor are there any necessary truths—at least not if we identify the latter with absolute and unrevisable truths, or if we follow the positivists in equating statements that express necessary truths with statements that express analytic truths.

Another of Quine's influential proposals is that traditional epistemology should be rejected in favor of "naturalized epistemology." Since Quine takes the physical sciences to be his primary sources of knowledge of reality, what used to be *philosophical* questions of ontology and epistemology become *scientific* questions. So, just as we turn to science in order to answer the question

"What is there in the world?" we turn to science to answer the question: "What do we know, and how do we know it?" *What* we know is covered by scientific knowledge; as for *how* we know what we know, we also look to science: in place of the standard epistemological questions, Quine asks: "How have we acquired science on the basis of our sensory stimulations?" Science gives us a picture of ourselves interacting with nature: we are biological beings whose surfaces are bombarded with a variety of stimulations. Quine accepts this picture and ponders the relationship between these stimulations and our conceptual scheme. He argues that this is largely a question of psychology and language learning: our basic categories come with our mother tongue, and to understand the relationship between these categories—that of "object," for example—we must study how we come to master language in response to sensory stimulations.

So for Quine, epistemology is simply science directed onto itself. Is it not circular to use science itself to establish the epistemological foundation of scientific theory? Quine does not deny that there is a circularity here. He defends it by pointing to the lack of any viable alternative. Traditional epistemology sought to *justify* our beliefs in two ways: the rationalists by appeal to necessary truths of reason, empiricists by showing how our beliefs can be deduced from experience. Quine rejects both approaches. There are no necessary or analytic truths of the sort the rationalist needs—this conclusion follows from the network model. He rejects empiricism because it is impossible to *deduce* any empirical proposition from an observational base. This follows from the network model too, more specifically the holistic thesis that sentences face the tribunal of sense collectively, not singly. If our belief-system cannot be grounded in either of the two traditional ways, we have no option but to accept it as largely true, and to critically assess it from this perspective:

I see philosophy not as an *a priori* propaedeutic or groundwork for science, but as continuous with science. I see philosophy and science as in the same boat—a boat which, to revert to Neurath's figure as I so often do, we can rebuild only at sea while staying afloat in it. There is no external vantage point, no first philosophy. (1969a, pp. 126–7)

Or again:

The naturalistic philosopher begins his reasoning within the inherited world theory as a going concern. He tentatively believes all of it, but believes also that some unidentified portions are wrong. He tries to improve, clarify, and understand the system from within. He is the busy sailor adrift on Neurath's raft. (1981, p. 72)

If we agree that physical science is the most impressive body of knowledge that we possess, and if we also accept that there is no possibility of a “first philosophy” that we can use to critically evaluate science from some higher or better vantage point, then we may feel tempted by Quine’s naturalized epistemology. Certainly many in analytic philosophy have been, in subsequent decades.

This stance leads to Quine’s realism. Despite believing that theories are underdetermined by data, Quine believes that the world really is as our best science says it is. Since much of science requires mathematics he is also prepared to commit himself to the existence of any entities that mathematics requires—sets, for instance. The objects that feature in the ontologies of our best theories are simply “posits,” intermediaries between sensation and sensation, introduced into a theory solely to improve its predictive and explanatory power, but they have a better claim on reality than anything else.

Quine’s realism is another respect in which he differs from Carnap. In his “Empiricism, Semantics and Ontology” (1950), the latter drew a distinction between *internal* and *external* questions. Internal questions are those that are addressed within a logico-linguistic system—the issues we seek to address in ordinary life and science are generally all internal questions. Carnap believed there that are many logico-linguistic systems to choose from, and that the main external question is “Which of these systems should we use?” Since rational argument is only possible when we have agreed on a language and a logic, it can only be deployed in addressing internal questions, hence there is no place for reason or rational argument in deciding how to answer *external* questions. Since realism and phenomenalism are simply alternative language-systems—at least for Carnap—there is no rational way of adjudicating between these systems. When doing science we may as well opt for a realistic framework, and Carnap believes we should because it is simpler than the available alternatives. But we should not conclude that realism is true, or that it is more justified than the alternatives—for again, justification only enters the picture within a logico-linguistic framework. The debate between realists and phenomenologists is a *metaphysical* one; as such—in Carnap’s eyes—it is not amenable to rational discussion, and so is a “pseudo-question,” rather than a genuine question. It follows that the realist’s claim that the objects posited by our scientific theories really exist is also metaphysical, and not to be taken seriously.

Quine will have none of this, and he rejects the distinction between internal and external question. There can only be a sharp distinction between logico-linguistic frameworks and empirical theories if there is a sharp distinction between analytic and nonanalytic truths, which of course Quine denies. Given the underdetermination thesis, there is—potentially at least—multiplicity at the level of scientific theories, and in deciding between the alternatives we will

inevitably be guided by pragmatic considerations: we are looking for a theory that *works*, one which is empirically successful.³ But once we have settled on our best theory, we should take it seriously and literally—and accept that everything the theory says exists does exist.

There may be no “first philosophy” for Quine, but this does not mean philosophy lacks a distinctive role. The proper task for the philosopher is the clarification of scientific discourse and scientific theories. Quine argues that as things currently stand, the best way of doing this is by formulating our theories in the language of first-order (predicate) logic, a task he calls “regimentation.” Under Quine’s preferred mode of regimentation Russell’s theory of descriptions is to be applied to all definite descriptions—hence the latter will thus be eliminated in favor of sentences that contain only quantifiers and bound variables. Quine goes further and proposes eliminating even simple proper names such as “Pegasus”: these can be replaced with suitably selected *predicates*, such as “is-identical-with-Pegasus” or “pegasizes.” So, for example, “Pegasus is a horse” becomes:

$$(\exists x)(\forall y)[(y \text{ is-identical-with-Pegasus iff } x = y) \ \& \ x \text{ is a horse}]$$

The objects that make quantified sentences such as these true are for Quine the objects to whose existence we are ontologically committed. Or as he puts it in his influential “On What There Is”: any theory “is committed to those and only those entities to which the bound variables of the theory must be capable of referring to in order that the affirmations made in the theory are true” (1953, p. 15).

The Quinean criterion of ontological commitment leaves open just one *mode* of existence: to exist is to be the value of a bound variable in a quantified sentence. The idealists, it will be recalled, held that there are *degrees* of existence. Not in Quine’s scheme of things. A no less striking consequence of Quine’s criterion is that objects of seemingly very different types all exist in the same way. The quantifiers in science—or so Quine maintains—range over both ordinary material objects (planets, chairs, etc.) and abstract nonspatio-temporal entities such as sets, numbers, and functions. For Quine both kinds of object exist in precisely the same clear and unambiguous manner.

With the outbreak of the Second World War Quine spent four years in US Naval Intelligence. He spent much of his time there attempting to decipher messages sent to and from German submarines; the messages were encoded by the (now famous) Enigma machines:

The Germans had a replica Enigma breaking complicated ciphers. Each day they had a different setting on the machine. We had to get it the hard way, by intercepting a message from a submarine that gave direction

finders. We would know, say, from the preceding day's message he had been sent on a refuelling rendezvous, so a good guess was that some word would be "refuelling." Then if our men could fit the word, they could get the setting for the whole.⁴

It is by no means impossible that all this time spent on decoding contributed to one of Quine's most discussed theses: the *indeterminacy of translation*. His most famous book, *Word and Object* (1960) opens with these lines:

Language is a social art. In acquiring it we have to depend entirely on intersubjectively available cues as to what to say and when. Hence there is no justification for collating linguistic meanings, unless in terms of men's dispositions to respond overtly to socially observable stimulations.

Although this may sound plausible and innocuous, in Quine's hands this constraint on meaning has dramatic consequences, as becomes clear in Chapter 2, where he develops a thought experiment in considerable detail.

We are invited to imagine a field linguist setting out to learn the previously untranslated language of a jungle tribe. Since the language of this tribe is quite unlike any other language ever studied, the linguist's task—"radical translation" as Quine calls it—is not going to be an easy one.⁵ It will involve developing a "translation manual" from the jungle language to English. This will comprise a dictionary, stating which native words correspond to English words, and a grammar: in short, it will be the kind of book one can buy to learn any other foreign language, such as French or Polish. It is not too difficult to envisage how the linguist might make a start. After establishing himself as a trustworthy friend, he would need to observe the tribe closely, noting the words and sentences they use in particular situations. Quine envisages a rabbit rushing by, a native pointing at it and shouting "Gavagai!" He goes on to suggest what might happen next: "The linguist notes the native's utterance of 'Gavagai' where he, in the native's position, might have said 'Rabbit', and looks to natives for approval. Encouraged, he tentatively adopts 'Rabbit' as translation" (1960, p. 42). In these conditions, although the linguist would inevitably note "Gavagai = rabbit?" in his notebook, he would await a good deal of further confirmation of this hypothesis before rubbing out the "?". But it is reasonable to expect that in due course the confirmation would come—assuming of course that the translation is correct.

And so it goes on: the linguist's task may be a difficult one, but, Quine suggests, there is no reason to believe that it cannot be completed. After all, did we not all start off as radical translators when attempting to learn our mother tongues as infants? The linguist will know that his job is done when his translation manual allows him to converse fluently with the natives with perfect

ease, about all manner of topics—for example, he is able to learn about the tribe's fishing and farming practices from talking with fishermen and farmers, its history, religion, and metaphysical beliefs from the tribal elders.

Quine now argues that something along the lines of the following scenario is possible. Let us suppose that, unbeknownst to our field linguist, a second field linguist has also been attempting to crack the jungle language—although arriving at roughly the same time, this competitor stayed in a different village, and so the two did not meet. This second linguist also succeeds in putting together a viable translation manual. He too can converse with the natives about everything under the sun.

Given that both linguists can talk to the natives with equal fluency, it is natural to suppose that the contents of each of their translation manuals will necessarily be very similar. But Quine argues that it is perfectly possible that the manuals will be very different, that the English equivalents they offer for many of the jungle-words will be radically divergent. Furthermore, even if the linguists decide to return to the jungle in an attempt discover which manual is correct, it is perfectly possible that they will fail: that even after extended conversations with the natives—carried out with the express purpose of determining which translation manual is correct—both manuals will still seem equally good.

Last but by no means least, Quine tells us that indeterminacy of translation “begins at home.” When as young children we first learn our native languages we are in exactly the same position with respect to our linguistic community as our field linguist was to the tribe. We both begin with no understanding of the language, and gradually arrive at one on the basis of observing speakers using their language. The empirical evidence available to the child and a radical field linguist is exactly the same. If indeterminacy is unavoidable in the jungle, it is likewise unavoidable at home. We naturally assume that when we use ordinary words such as “rabbit,” “table,” “chair,” “moon,” we are justified in assuming, as we do, that we are all referring to the same things. But according to Quine this natural assumption is mistaken.

As for why we should believe that translation *is* indeterminate in this manner, Quine offers a variety of arguments, which we cannot explore in any detail here, but all of which have provoked much discussion. An “argument from above” relies on the widely accepted general thesis that scientific theories are underdetermined by data, that for any given collection of data there will always be many different theories that explain it equally effectively. Since a translation manual is, in effect, a *theory* as to the meaning of words in the target language, theories about translation too will be underdetermined by data. There is also an “argument from below.” Although it is natural for the linguist to assume “Gavagai!” means *rabbit*, there is any number of alternative translations—for example, *rabbit-complement*, *fusion of rabbit-parts*, *rabbit-slice*, instantiation of *rabbit-hood*—which are found whenever and wherever

rabbits are found, and which cannot be ruled out.⁶ Quine argues that provided *other* expressions in the native language are translated in accord with, say, the hypothesis that “Gavagai!” means “There is an instantiation of rabbit-hood” rather than “There is a rabbit,” the resulting translation manuals will be equally effective at communicating with the natives themselves, and hence empirically indistinguishable.

If two competing translation schemes are empirically equivalent they are equally valid from a scientific perspective, and so Quine concludes that we should view them as being equally valid *tout court*. He is thus led to the conclusion that the indeterminacy of translation thesis is not just an epistemological matter, it also carries ontological import: when confronted with empirically equivalent translation schemes, it is not simply a matter of our not being able to discover which of the schemes is the correct one, according to Quine there is *no fact of the matter* as to which is correct.

. . . adopt for now my fully realistic attitude towards electrons and muons and curved space-time, thus falling in with the current theory of the world, despite knowing that it is methodologically under-determined. Consider, from this realistic point of view, the totality of truths of nature, known and unknown, observable and unobservable, past and future. The point about indeterminacy of translation is that it withstands even all this truth, the whole truth about nature . . . This is what I mean by saying that, where indeterminacy applies . . . there is no fact of the matter even to *within* the acknowledged under-determination of a theory of nature. (1969b, p. 303)

Bearing in mind that indeterminacy begins at home, there is thus no fact as to the matter with regard to whether any of *us* mean *rabbit* or *rabbit-complement* or *rabbit-slice*, etc. when we talk and think about rabbits.

All this gives rise to a number of pressing questions. If there are no facts of the matter with regard to meaning and reference, what remains of Quine’s commitment to scientific realism? Is it still possible for us to believe that the entities posited by our best theories really exist? If there are no facts about meaning, what becomes of Quine’s own arguments for the indeterminacy thesis? How can they fail to be self-undermining? Needless to say, Quine has responses to these questions, but we cannot explore them here. These issues aside, there remains the question of precisely what it is that Quine thinks he has achieved by his indeterminacy arguments. One of his targets is on clear display as in this passage from his essay “Ontological Relativity”:

Uncritical semantics is the myth of a museum in which the exhibits are meanings and the words are labels. To switch languages is to change

the labels. Now the naturalist's primary objection to this view is not an objection to meanings on account of their being mental entities, though that could be objection enough. The primary objection persists even if we take the labelled exhibits not as mental ideas but as Platonic ideas or even as the denoted concrete objects. Semantics is vitiated by a pernicious mentalism as long as we regard a man's semantics as somehow determinate in his mind beyond what might be implicit in his dispositions to overt behaviour. (1969a, p. 27)

For Locke, understanding a language is, at bottom, a matter of associating the correct internal "ideas" with the correct words; these mental (image-like) entities are the primary locus of meaning. Frege and Russell both held that meaning resides in *propositions*: abstract Platonic entities that we mentally apprehend or recognize. On both sorts of view, there is a definite fact of the matter as to whether we mean "That's a *rabbit*" or "That's a *collection of rabbit-parts*" simply because these are two very different propositions—or in the Lockean case, different combinations of ideas. On either of these views, there is room for there to be more to the meaning that we attach to words than is recoverable from how we use words in observable circumstances: what we mean by the use of a particular sentence on a certain occasion is determined by something that is mentally apprehended. For Quine, these ways of thinking about language and meaning belong to the realm of prescientific mythology. Propositions and meanings, thus construed, have no role to play in a scientifically informed view of language and communication.

There is a further point that should not be overlooked. Recall the alternative translations that Quine proposes for "Gavagai!": brief temporal rabbit-phase, instantiation of rabbit-hood, fusion of rabbit-parts, and so forth. These all correspond to different *metaphysical* conceptions of compound material objects—indeed, there are philosophers who, in recent years, have advocated abandoning our ordinary way of thinking of such objects in favor of one or other of these alternative modes of conceptualization. Since this choice of example is unlikely to be coincidental, it is by no means improbable that traditional *metaphysics* is something else that Quine hoped to undermine with his indeterminacy theses.

11

Oxford and Ordinary
Language*Barry Dainton*

The influence of the later Wittgenstein's work was initially felt at Cambridge, but it was not long before it spread. Although the *Investigations* would not be published until 1953, reports of Wittgenstein's lectures and typescripts of his writings soon found themselves in circulation from the early 1930s, and one center of philosophical activity where they made a deep impact was Oxford, which would soon become a leading center of analytical philosophy in its own right.¹ Hacker writes:

In 1939 philosophy at Oxford was poised for a renaissance. This was delayed by the war. However the creative abilities damned by six years away from academia then flowed all the more powerfully when university life revived. The younger generation returned to philosophical work matured by their years at war, and post-war Oxford saw a spectacular philosophical flowering. (1996c, p. 148)

If Gilbert Ryle and J. L. Austin were the leading lights in the early part of the 1945–70 period in Oxford philosophy that Hacker is here referring to, Paul Grice and Peter Strawson took over that mantle later on, with David Pears, Stuart Hampshire, E. Anscombe, R. M. Hare, W. D. Hart, and Geoffrey Warnock lending assistance.

The label “ordinary language” is associated with Oxford philosophy in this period. As commonly characterized—and ignoring the various minor differences in doctrine and emphasis among the protagonists—the ordinary language school took the view that philosophical problems are largely a product of the misuse, or misconstrual, of language, and that these problems should be solved—or dissolved—through the assembling of reminders as to the way in which the relevant words are used in ordinary life. Philosophical analysis is thus *not* primarily a matter of revealing the hidden logical forms lying beneath ordinary language—as the early Wittgenstein and Russell believed. The Oxford linguistic philosophers did agree, however, with the later Wittgenstein's view

that meaning is use (at least as a first approximation), and that philosophical investigations should be conducted on a piecemeal, case-by-case basis; systematic theories of language or meaning are not needed, or appropriate. They agreed with Moore that our ordinary everyday beliefs are foundational: that we can perceive medium-sized material bodies, such as hands and houses, is not to be doubted.

While this characterization of the Oxford school's doctrines is more than a caricature—it corresponds reasonably well with the work of Ryle and Austin—it does less than full justice to the work of Grice and Strawson. For although the latter pair started their careers as ordinary language practitioners, as we shall see, it would not be long before they moved in different directions. Indeed, Austin's own later work—before his early death—was moving in the direction of systematic theorizing.²

Austin published comparatively little during his lifetime, but this did not prevent him from exerting a powerful influence on Oxford philosophy, particularly during the 1950s—he died in 1960, aged only 48. During the 1950s he held regular Saturday morning classes for the junior (nonprofessorial) fellows, and these proved to be an effective way of conveying and developing his distinctive approach to philosophical issues.³ Austin's seven published papers include "Other Minds" (1946), "Ifs and Cans" (1956), and "A Plea for Excuses" (1956). His two books, *Sense and Sensibilia* (1962) and *How to do Things with Words* (1962), were both posthumous publications, based on lecture notes.

In "A Plea for Excuses" Austin offers the following by way of a defense of the ordinary language approach:

First, words are our tools, and, as a minimum, we should use clean tools: we should know what we mean and what we do not, and we must forearm ourselves against the traps that language sets us. Secondly, words are not (except in their own little corner) facts or things: we need therefore to prise them from the world, to hold them apart from and against it, so that we can realize their inadequacies and arbitrariness, and can relook at the world without blinkers. Thirdly, and more hopefully, our common stock of words embodies all the distinctions men have found worth drawing, and the connexions they have found worth marking, in the lifetimes of many generations: these surely are likely to be more numerous, more sound, since they have stood up to the long test of the survival of the fittest, and more subtle, at least in all ordinary and reasonably practical matters, than any that you or I are likely to think up in our arm-chairs of an afternoon—the most favoured alternative method. (1956–7, pp. 7–8)

He moves on to address a frequently made objection to linguistic philosophy: do usages not differ? Do different people not say different things in the same

sorts of circumstance? Austin claims that this happens less often than one might think, and if it does happen in respect of an imaginary case, describing the case in more detail will usually lead to agreement:

As practice in learning to handle this bogey . . . we could scarcely hope for a more promising exercise than the study of excuses. Here, surely, is just the sort of situation in which people will say “almost anything,” because they are so flurried, or so anxious to get off. “It was a mistake,” “It was an accident”—how readily these can *appear* indifferent, and even be used together. Yet, a story or two, and everybody will not merely agree that they are completely different, but even discover for himself what the difference is and what each means. (ibid., pp. 10–11)

Austin provides an illustration of what he has in mind in a famous footnote:

You have a donkey, so have I, and they graze in the same field. The day comes when I conceive a dislike for mine. I go to shoot it, draw a bead on it, fire: the brute falls in its tracks. I inspect the victim, and find to my horror that it is your donkey. I appear on your doorstep with the remains and say—what? “I say, old sport, I’m awfully sorry, &c, I’ve shot your donkey by *accident*”? Or “by mistake”? Then again, I go to shoot my donkey as before, draw a bead on it, fire—but as I do so, the beasts move, and to my horror yours falls. Again the scene on the doorstep—what do I say? “By mistake”? Or “by accident”? (ibid., p. 11)

After only a little reflection it is obvious (to most of us) that there is indeed a difference, and an important one, between accidents and mistakes.

Of Austin’s two posthumous books, *How to Do Things with Words* is probably the more influential, for it would give rise to the new discipline of “speech act theory,” to which a number of other philosophers—for example, Strawson, Grice, Searle, Recanati, Harnish, and Bach—would make important contributions in the years to come. Austin starts off by criticizing the widespread assumption among philosophers writing on language that the purpose of sentences can only be to describe states of affairs or state facts, truly or falsely. Fact-stating is certainly one of the things we do with words—it is one type of “speech act”—but there are others. Austin draws our attention to “explicit performative utterances” such as “You’re fired,” “I nominate . . .” or “You are hereby sentenced. . . .” These are of interest because the utterances of these sentences, in the appropriate contexts (e.g. a court of law), are *themselves* the very acts of the kind specified by the verb (a firing, a nomination, a sentencing). We thus need to recognize that a distinction exists between saying X, and the *effect achieved* by saying X.

Austin identifies three aspects of speech acts: (a) the act *of* saying something, (b) what one does *in* saying it, and (c) what one does *by* saying it; he labels these the “locutionary,” the “illocutionary,” and the “perlocutionary” respectively. By way of illustration, suppose a pilot makes this onboard announcement: “The plane will be landing in fifteen minutes.” By so doing he performs the *locutionary* act of saying that the plane will land in 15 minutes time. The pilot is also performing the *illocutionary* act of telling the plane’s passengers that the plane will land soon, and (no doubt) implying that they prepare themselves accordingly (e.g. by ordering their final drinks, or visiting the WC). If this illocutionary act produces the desired effect—if it “produces uptake”—the passengers will understand what the announcement says. However, the pilot is also performing a *perlocutionary* act: as well as wanting to be understood, he wants his utterance to produce in the passengers the belief that the plane will soon be landing, and to respond appropriately (by ordering their last drinks). The three acts—or act aspects—are performed simultaneously, with the one linguistic act, the one utterance. Austin’s terminology is now standard.

In *Sense and Sensibilia* Austin’s aims are more negative: the book contains a barrage of criticism directed against the doctrine—advocated by positivists such as Ayer—that the objects of immediate perceptual acquaintance are sense-data, and that statements about sense-data are epistemically foundational. Here is Austin taking issue with Ayer’s claim that statements about appearances (or sense-data) are evidence for statements about physical objects:

... it is not the case, as this doctrine implies, that whenever a “material object” statement is made, the speaker must have or could produce evidence for it. This may sound plausible enough; but it involves a gross misuse of the notion of “evidence.” The situation in which I would properly be said to have *evidence* for the statement that some animal is a pig is that, for example, in which the beast itself is not actually on view, but I can see plenty of pig-like marks on the ground outside its retreat. If I find a few buckets of pig-food, that’s a bit more evidence, and the noises and smell may provide better evidence still. But if the animal then emerges and stands there plainly in view, there is no longer any question of collecting evidence: its coming into view doesn’t provide me with more *evidence* that it’s a pig, I can now just *see* that it is, the question is settled. (1962, p. 115)

The work is by no means entirely negative. Austin also provides valuable discussions of many concepts associated with perception, for example, the differences between illusions, deceptions, and hallucinations, between “veridical,” “direct,” and “indirect” perception, between how things *look*, *appear*, and *seem*, and so forth.

Gilbert Ryle was an early convert to the ordinary language cause. Ryle was appointed to the Oxford Waynflete Chair in 1945, and in 1947 he took over (from Moore) the editorship of *Mind*, which he would retain until 1971. Ryle published what would become his most famous book, *The Concept of Mind*, in 1949. Here he attacks the “Official Doctrine,” the view—associated with Descartes—that the mental and the physical belong to distinct realms, with the physical existing in space and subject to mechanistic laws, with the mental *not* residing in physical space, and not subject to physical laws. Ryle argues that a proper understanding of mental concepts—concepts such as *knowing*, *imagining*, *learning*, *hoping*, *wanting*, *feeling*, *feeling a pain*, *doing voluntarily*, *doing deliberately*, *remembering*, *perceiving*—reveals the absurdity of the Official Doctrine, for it soon becomes clear that the mental and the physical are far more intimately interrelated than this doctrine can accommodate. In real life there are behavioral criteria for knowing something, feeling a pain, remembering something, criteria that in ordinary circumstances settle beyond all reasonable doubt whether someone really does (say) remember doing something on a particular occasion, or has learned how to perform mental arithmetic, or believes this or that. Indeed, for Ryle in an important sense there are not really “minds” at all:

One of the central negative motives of this book is to show that “mental” does not denote a status, such that one can sensibly ask of a given thing or event whether it is mental or physical, “in the mind” or “in the outside world.” To talk of a person’s mind is not to talk of a repository which is permitted to house objects that something called “the physical world” is forbidden to house; it is to talk of the person’s abilities, liabilities and inclinations to do and undergo certain sorts of things, and of the doing and undergoing of these things in the ordinary world. Indeed, it makes no sense to speak as if there could be two or eleven worlds. Nothing but confusion is achieved by labelling worlds after particular avocations. Even the solemn phrase “the physical world” is as philosophically pointless as would be phrase “the numismatic world,” “the habershashery world” or “the botanical world.” (1949, p. 199)

Several of the logical positivists, notably Carnap (1932–3) and Hempel (1949), advocated a reductive form of behaviorism. Ryle is no reductionist, and his motivations are very different—he is not in the least concerned with reconciling the language of psychology with that of physics—but he is leaning in the same general direction.

The methodology Ryle employs in *The Concept of Mind* is one he had been elaborating and refining over the previous decade and a half. In his 1932 paper

"Systematically Misleading Expressions" he clearly set out his stand from the outset, the paper beginning thus:

Philosophical arguments have always largely, if not entirely, consisted in attempts to thrash out "what it means to say so and so". . . . Sometimes philosophers say that they are analysing or clarifying the "concepts" which are embodied in the "judgments" of the plain man or of the scientist, historian, artist or who-not. But this seems to be only a gaseous way of saying that they are trying to discover what is meant by the general terms contained in the sentences which they pronounce or write. (p. 139)

Among the many who spoke in terms of concepts and judgments were, of course, Russell and Moore. Ryle goes on to note, however, that this "whole procedure is very odd." The intelligent, ordinary users of the expressions being analyzed know perfectly well what they mean, and do not need to wait for the philosophers to tell them—a point the Wittgenstein was also starting to argue at around this time. As for philosophers, they too must understand the expressions in question, for if they did not they would not know what it was they were analyzing. So how is informative analysis possible? It seems "that if an expression can be understood, then it is already known in that understanding what the expression means. So there is no darkness present and no illumination required or possible" (p. 140).

Ryle is here drawing our attention to what became known as "the paradox of analysis"—see Black (1944)—and he goes on to provide a solution. Certain expressions have the distinctive property of being misleading in a particular way: they are *systematically misleading*. Expressions in this category are perfectly well-understood by the ordinary people who use them, however they are "couched in grammatical or syntactical forms which are in a demonstrable way *improper* to the states of affairs which they record" (p. 142) and it is these that are liable to give rise to philosophical mischief. As his first example, Ryle compares "I am sleepy" with "Satan does not exist" and "Carnivorous cows do not exist." If we rely on nothing but ordinary grammar, then "Satan" and "carnivorous cows" look as though they are referring expressions, in the manner of "I." But if we take them to function in the same way, then "Satan does not exist" will have to be *about* something, there will have to be something for "Satan" to refer to, even though Satan does not exist. Some have introduced nonactual objects, or mental entities (an "idea" of Satan, say) to fill this role, but as Ryle notes, since we also say things like "round squares do not exist" and "real nonentities do not exist," if we follow this course we are "bound to fill the realm of subsistents or the realm of ideas with walking self-contradictions" (p. 144), so an alternative analysis is needed. A better option is to ignore the grammar, and take "carnivorous cows do not exist" to mean "no

cows are carnivorous" or no carnivorous beasts are cows." We can treat "Satan does not exist" in a similar way. Rather than taking "Satan" to be a subject expression that refers to something, we construe it as a predicative expression, one that asserts that something has certain properties, "nothing is both devilish and alone in being devilish" or "nothing is both devilish and called 'Satan.'" Ryle suggests that "carnivorous cows do not exist" and expressions with a similar form are *systematically* misleading, but they are neither false nor senseless: they mean precisely what the proposed paraphrases mean.

What Ryle is recommending here, of course, is much the same as the approach advocated by Russell in "On Denoting." However, unlike Russell, Ryle's analysis is conducted in ordinary language: he does not feel it necessary to employ the resources of an artificial language when expounding his analyses.⁴ Ryle's case against "concepts" and "judgments" and the like rests on similar considerations. Consider "Jones hates the thought of going to hospital" or "the idea of having a holiday has just occurred to me." If we take the surface grammar to be a guide to the ontological concomitants of such claims, then we are going to have to recognize Platonic thoughts, Lockean ideas or somesuch. But, Ryle argues, a little paraphrase here and there reveals that these metaphysical moves are not in fact needed. We can reformulate the first as "Jones feels distressed when he thinks of what he will undergo if he goes to hospital" and the second as "I have just been thinking I will take a holiday." As soon as we make these moves, any temptation to introduce *thoughts* or *ideas* vanishes like the morning mist.

I suspect that all the mistaken doctrines of concepts, ideas, terms, judgments, objective propositions, contents, objectives and the like derive from the same fallacy, namely, that there must be *something* referred to by such expressions as "the meaning of the word (phrase or sentence) 'x' on all fours with the policeman who really is referred to by the descriptive phrase 'our village policeman is fond of football.'" (p. 162)

Those who have felt obliged to endorse a Platonistic view of propositions or concepts might well be skeptical whether such momentous ontological savings can be made by the comparative simple paraphrastic manoeuvres of the sort that Ryle wields against them. That such simple measures *could* have such momentous consequences is precisely what filled the ordinary language pioneers with at times near-missionary zeal. In a 1971 interview Strawson describes the approach taken by Austin and Ryle in the early days of the ordinary language movement:

There was something in common to their methods at that time, though the style was very different. They both gave careful attention to what

could, or couldn't, be naturally or non-absurdly *said*; and also to the circumstances in which we could or couldn't naturally say such and such a thing. And this method, for reasons which seemed obvious enough when they were pointed out, was a very fruitful source of philosophical data. . . .

He then explains why it generated the excitement that it did:

The reasons why the thing was exciting were really two. On the one hand, in the face of this refined examination of actual linguistic practice, a lot of traditional philosophical theorizing began to look extraordinarily crude, like an assemblage of huge, crude mistakes. And it was, of course, extremely exhilarating to see these huge and imposing edifices of thought just crumbling away, or tumbling down, to the tune of this fairly modest sort of piping. And then, on the other hand, there was something else, something more positive: the sense of discovery of the fine, subtle texture of our actual thinking, of our actual conceptual and linguistic equipment. (Magee 1971, p. 116)

Strawson started to establish his own reputation in 1950, when in different articles he attacked both Austin and Russell. In a symposium with Austin on the topic of truth, Austin defended a version of the correspondence theory (between propositions and facts) and Strawson cast doubt on the intelligibility of talk of correspondence in this context. Strawson had begun to elaborate an alternative "performative" conception of truth a year earlier (1949). He notes that there is a close relationship between saying that it is true that S, and simply saying S. For example, consider:

It is true that the earth is round.

The proposition that the earth is round is true.

That's true (said in response to someone's saying that the earth is round).

Strawson contends that anyone who makes any of these claims is saying the same thing as someone who says "The earth is round." In which case, the addition of the predicate "is true" in the sentences above adds nothing to what they are being used to assert. For Strawson, truth is not a *property* that we ascribe to propositions. Rather, when we say that a sentence S is true, we are using "true" to perform a speech act of, in this case, endorsing S.

In his most famous article, "On Referring" (1950), Strawson took issue with Russell's theory of descriptions. As we have seen, Russell analyzed sentences such as "The present king of France is bald" as "There is a unique

person who is the present king of France, and that person is bald"—an analysis that leads to the sentence's being false in a straightforward way. It is by no means obvious, however, that the *analysans* and *analysandum* here have precisely the same meaning. Nor is it self-evident, Strawson suggests, that an ordinary person, when confronted with this sentence, would agree that it is clearly and straightforwardly false. Russell felt obliged to offer an analysis along the lines he did because he took the view that all meaningful (and purportedly fact-stating) statements must be either true or false. Strawson questions the necessity of this. In normal discourse, an utterance of a sentence such as "The present king of France is bald" is accompanied by a certain *presupposition*:, namely that there exists something to which the description applies, a current French monarch.⁵ What is presupposed in such cases is not part of what the speaker asserts when using a sentence of the form "The *F* is *G*," but it is there as a background assumption. Accordingly, if, as in this case, this presupposition is not satisfied, the rules of the game have been infringed, and the utterer of the sentence fails to make a true *or* a false statement—indeed, the question of truth or falsity does not even arise.

Strawson published his first book, *Introduction to Logical Theory*, in 1952. The book is concerned with the nature and limitations of logic, and one of the more influential discussions in the book concerns the extent to which the ordinary language logical connectives, such as "and," "or," and "if . . . then . . ." can be equated with the truth-functional connectives &, v and \rightarrow of formal logic. Strawson's second book, *Individuals* appeared in 1959. It is generally considered to be his most significant.

In the first lines of his introduction to *Individuals* Strawson contrasts two approaches to metaphysics: "Metaphysics has often been revisionary, and less often descriptive. Descriptive metaphysics is content to describe the actual structure of the world, revisionary metaphysics is concerned to produce a better structure." Strawson concedes that no metaphysician has ever been solely revisionary or descriptive, in intention or effect, but suggests that Descartes, Leibniz, and Berkeley are broadly speaking revisionary, whereas Aristotle and Kant are descriptive. If the best works in the tradition of revisionary metaphysics have enduring interest, thanks to the "intensity of their partial vision," it is the less heralded descriptive metaphysics that, Strawson suggests, has the real and greater value. We are told that "there is a massive central core of human thinking which has no history—or none recorded in histories of thought; there are categories and concepts which, in their most fundamental character, change not at all," and it is the task of descriptive metaphysics to bring into clear view the conceptual structures and interrelationships that constitute this core system. Strawson anticipates an objection: how will this

descriptive metaphysics differ from logical or conceptual analysis of the more familiar kind? It does not differ in “kind of intention, but only in scope and generality.” He continues:

Up to a point, the reliance upon a close examination of the actual use of words is the best, and indeed the only sure, way in philosophy. But the discriminations we can make, and the connexions we can establish, in this way, are not general enough and not far-reaching enough to meet the full metaphysical demand for understanding. For when we ask how we use this or that expression, our answers, however revealing at a certain level, are apt to assume, and not to expose, those general elements of structure which the metaphysician wants revealed. The structure he seeks does not readily display itself on the surface of language, but lies submerged. He must abandon his only sure guide when the guide cannot take him as far as he wishes to go.

Metaphysics, of one sort at least, is back on the philosophical agenda.

Individuals is divided into two parts. As Strawson himself puts it the “first part aims at establishing the central position which material bodies and persons occupy among particulars in general,” whereas in “the second part of the book the aim is to establish and explain the connexions between the idea of a particular in general and that of an object of reference or logical subject.” It is the four chapters of the first part that have occasioned most debate subsequently.

In chapter one, “Bodies,” Strawson’s focus is on our actual system of concepts, and he explores the issue of whether there is a category of thing that we can refer to independently of our being able to refer to things of a different category. After an elaborate argument, material bodies emerge as “referentially basic” particulars. The objects to which we ordinarily exist in the same space and time as we do, but the spatiotemporal framework, as Strawson conceives it, is a system of relations that are themselves grounded in our ability to identify and reidentify material bodies through space and over time. Chapter 2 is intriguingly entitled “Sounds.” Is the concept of a persisting mind-independent particular essentially bound up with the spatio-temporal world of material bodies? Or can the notion of *objectivity* find application in a very different conceptual scheme? To investigate these issues Strawson considers what life might be like for subjects who inhabit an entirely *non-spatial* universe. Visual experience (at least of the sort we have) is inherently spatial, but auditory experience is not—or at least, it is not essentially spatial. So Strawson invites us to imagine the life of a subject who *only* has auditory experience, and poses this question: “Could a being whose experience is purely auditory have a conceptual scheme which provided for objective

particulars?" (1959, p. 66). Strawson again embarks on a complex argumentative journey, and arrives at this conclusion:

. . . to have a conceptual scheme in which a distinction is made between oneself or one's states and auditory items which are not states of oneself, is to have a conceptual scheme in which the existence of auditory items is *logically* independent of the existence of one's states or of oneself. Thus it is to have a conceptual scheme in which it is logically possible that such items should exist whether or not they were being observed, and hence should continue to exist through an interval during which they were not being observed. So it seems that it must be the case that there could be reidentifiable particulars in a purely auditory world if the conditions of a non-solipsistic consciousness could be fulfilled for such a world. Now it might further be said that it makes no sense to say that there logically could be reidentifiable particulars in a purely auditory world, unless criteria for reidentification can be framed or devised in purely auditory terms. And if this is correct, as it seems to be, we have the conclusion that the conditions of a non-solipsistic consciousness can be satisfied in such a world only if we can describe in purely auditory terms criteria for identification of sound-particulars. (1959, pp. 72–3)

As for whether the experience of a subject in the purely auditory universe *could* be such as to permit the concept of reidentifiable particulars to find application, Strawson is cautiously optimistic. We are able to reidentify objects here in our own world by virtue of being able to move away from them and then return to revisit them. Is there an analogue of this in an auditory universe? Strawson argues that there might be, if the circumstances are right—if the sounds behave in certain ways. One promising possibility involves a *master sound*: a sound that is always present, but which is of varying pitch, and along which one can move at will. Being stationed "at" particular pitches of the master sound constitutes an auditory analogue of occupying a particular spatial location, and this is enough to allow us—or so Strawson argues—to make sense of the notion that reidentifiable particulars exist in the auditory universe.⁶

Of the remaining chapters of *Individuals* it was the third, "Persons," which has stimulated most discussion. Here Strawson argues that in our ordinary thought we conceive of ourselves and others as essentially possessing two aspects: the mental and the physical. We are not, as Descartes held, wholly mental and nonspatial. Strawson here relies on the claim that we can ascribe experiences to ourselves only if we are able to ascribe them to others, and that we would not be able to do this if other subjects were wholly mental (and nonspatial) entities. Strawson goes on to draw an epistemological conclusion: our

ability to ascribe experiences to ourselves presupposes an ability to ascribe experiences to others. If Strawson is right there cannot be a genuine problem of other minds.

One of the distinguishing features of *Individuals* is the attention paid to distinctively Kantian themes—such as the conditions of possibility of identifying certain items—and in his next major work, *The Bounds of Sense* (1966), Strawson takes on Kant himself. His aim: to find those parts of Kant's *Critique of Pure Reason* that are worth preserving, and those that should be jettisoned. Here is Strawson himself talking about the first *Critique* and his attitude toward it:

There is, in the work a body of doctrine about the necessary general structure of experience . . . about the limits of what we can make truly intelligible to ourselves as a possible structure for our own experience. Now, this body of doctrine, though not acceptable in all respects, is in its general outline and many substantial points, I think correct. But it's surrounded by, and in Kant's own view it's dependent on, another, second body of doctrine, probably that by which he's best known. And this is the doctrine that the nature of things as they really are, or as they are in themselves, is necessarily completely unknown to us—that the world as we know it, including our ordinary selves, is mere appearance. . . . Now all this second body of doctrine I take to be a kind of nonsense, though it has a certain appealingly dramatic and exciting quality, like most metaphysical nonsense. So I conceived my task to be that of extracting as it were the kernel of truth and sense—that's to say, the first body of doctrine—from the surrounding shell of falsehood and nonsense—that's to say, the second body of doctrine.⁷

Whether or not Kant is indeed best served by separating his transcendental idealism from his transcendental arguments is an issue that has engaged Kantian scholars and sympathizers ever since. What is not in doubt is the importance of Strawson's contribution to Kantian commentary.

Strawson broadened the horizons of Oxford philosophy by rehabilitating (one form of) metaphysics. Paul Grice did likewise, but by reflecting on the nature of meaning and language. His important paper "Meaning" (1957) begins modestly enough, by focusing on the way we actually use the word "meaning" in ordinary language. Grice finds two main senses. We often say things such as "Those clouds mean rain," "Those spots mean measles," or "The recent budget means that we'll have a hard year." What we mean by *this* use of "mean" is that one event is going to bring about (cause) another event. Grice calls this "natural meaning." This type of meaning is independent of any sort of person to person communication. But now consider: "Those three rings on the bell (of the bus) mean that the bus is full," or "That remark,

'Smith couldn't get on without his trouble and strife' meant that Smith found his wife indispensable." Here an action—in the first case the ringing of a bell, in the second the utterance of a remark—is performed with the intention of communicating something to an audience. Grice calls this *non-natural meaning*. Generally speaking, non-natural meaning involves an agent performing an action that communicates something to an audience. The action is often linguistic, but it need not be.

Grice argues that communication consists of a hearer recognizing or inferring a speaker's communicative intentions. The latter usually have two components: the speaker intends that a particular act be recognized *as* an attempt to communicate, and they also intend to impart some particular piece of information. To see the relevance of this, compare the following:

- (a) Z shows Mr X a photo of Mr Y displaying undue familiarity with Mrs X.
- (b) Z draws a picture of Mr Y behaving in this manner and shows it to Mr X.

In the case of (a), Mr X's *recognition of Z's intention* to get him to believe that there is something between his wife and Mr Y is largely irrelevant to the production of this effect by the photo—after all, the same effect on his beliefs would have been produced if Mr X had found the photo lying around somewhere. In case (b) however, it will make a difference to the effect of Z's action whether or not X takes Z to be *intending* to inform him—that is, make him believe something—about Mrs X and not just to be doodling or producing a work of art. According to Grice, the difference is in *how* the belief transmission is induced. For genuine communication between utterer and audience, the recognition of the utterer's intention by the audience is crucial to the success of the attempted act of communication. More generally, for an agent *A* to non-naturally mean anything by *x*, *A* must intend by uttering *x* to induce a belief in an audience, and he must also intend his utterance to be recognized as so intended.

We have here an illuminating analysis of speakers-meaning, of what it is for a speaker to communicate something to someone else. Moreover, this analysis makes no use of the concept "meaning" or any other semantic notions: the key concepts used in the analysis are "intention" and "belief." It is thus both informative and noncircular. However, the analysis is limited to the meaning of specific communicative actions. Linguistic expressions have a "timeless" meaning, a meaning that is generally or conventionally understood, that is independent of specific speaker-hearer interactions, and that we learn when we learn a language. How can this be accommodated within the communicative action theory? Grice and his followers here appeal to a

process of fossilization through convention. Timeless expression-meaning is a conventional matter: it is a convention of English speakers to take "Ouch" as a sign that its utterer is in pain, similarly for more complex expressions such as "That hurts!" That certain sounds have the meaning they do is a matter of custom and practice, of how we traditionally and habitually use expressions.

In his earlier articles Grice does not go on to say much about how hearers go about working out what it is that a speaker wants them to understand, but he remedied this in later publications (e.g. 1975, 1989). There might not seem to be much of a difficulty here at all. Given that sentences of a language do have a conventional meaning, then surely if a speaker utters a sentence, provided the audience grasps that the speaker intends to convey whatever it is that the sentence means, no further work needs to be done. But in fact, as Grice shows, more has to be done, because even the most straightforward sentence is open to a wide range of different interpretations.

Grice begins by showing that we need to distinguish between what a sentence strictly and literally *says*, and what it *implies* or *conveys*; he calls the former "conventional meaning" and the latter "conversational implicature." Suppose two friends, A and B, are talking about a mutual friend C, who is now working in a bank. A asks B how C is getting on in his job, and B replies, "Oh, quite well, I think; he likes his colleagues, and he hasn't been sent to prison yet." At this point, A might well wonder what B was implying here: for in the context of this interchange, he is clearly implying more than he is saying: perhaps that C is the sort of person likely to yield to temptation, or that his colleagues in the bank are a bunch of crooks. It seems clear in this example that what B implied differed from what he *said*, which is just that C had not yet been sent to prison. It is also clear that what B implies might be false, since C might be honest and working in an honest bank, yet what he said still be true. Or, another famous example, suppose A is the student of B. A has applied for a job as a philosophy lecturer, and B has been asked to write a reference, saying what his opinion is of A's quality as a philosopher. B writes:

A has beautiful handwriting.

What this *says* may well be true. But in the context it conveys a lot more than this: it implies that B thinks A is not a good philosopher. How is this additional information conveyed? Clearly, if B thought A was any good, it would have been normal for him to say this; by not saying it, he conveys that he does not think it: if he did, he would have said so in the reference.

Grice tries to explain the underlying principles behind conversational implicatures. He suggested that conversations are typically *cooperative efforts*.

Talking and linguistic communication is a special case of purposive rational activity. We typically talk to another person with certain ends in view, ends we both share. Given these mutual goals, certain conversational moves are excluded as unsuitable in the context. Grice suggests that our normal exchanges are governed by this rule:

Cooperative Principle (CP): make your conversational contribution as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged.

He assumes here that a basic purpose in dialogue or conversation is the *maximally effective exchange of information*. This is not the only purpose, but it is an obviously important one: each side of a conversation has an interest in learning what the other side knows. Grice goes on to argue that certain conversational *maxims* followed from the CP, namely those of Quantity, Quality, Relation, and Manner:

- Quantity: make your contribution as informative as required for the current purpose of exchange; do not make your contribution more informative than is required.
- Quality: try to make your contribution one that is true; do not say what you believe to be false, do not say that for which you lack adequate evidence.
- Relation: be relevant, don't say things which are irrelevant to the course and purpose of the interchange.
- Manner: be perspicuous, avoid obscurity of expression, avoid ambiguity, be orderly, avoid unnecessary prolixity, and so on.

It is Grice's contention that in the vast majority of communicative interchanges we do tend to adhere to these principles, and this fact is explicable: it is in all our interests to do so, for the most part. There are, however, a number of ways in which a participant in a conversation may fail to fulfill a maxim. For example: (a) he may quietly and unostentatiously violate a maxim; in which case he will be liable to mislead; (b) he may make it clear that he is opting out of the CP; (c) he may be faced with a *clash*: for example, be unable to fulfill the maxim of quantity (be as informative as required) without violating the maxim of quality (have adequate evidence for what you say); (d) he may *flout* a maxim—blatantly fail to fulfill it.

In cases where the speaker *is* able to fulfill the maxim, and do so without violating another maxim, is not opting out, and not trying to mislead (or so the hearer concludes), the hearer is left with a problem: how can what he is saying be reconciled with his observing the overall CP? It is *this* type of situation

that typically gives rise to conversational implicature—and when this occurs a maxim is being *exploited*.

For example, in response to the prison remark, a hearer might reason as follows: (a) B seems to have flouted the maxim “be relevant” and/or maxims of the sort “be perspicuous”; (b) there’s no reason to think B has opted out of the CP; (c) assuming he is in fact conforming to the maxims, to make sense of what he says I have to assume he believes C to be dishonest; otherwise his remark does not make sense; (d) B knows I am capable of working this out; (e) Therefore, I take it that B intended to *imply* that C is dishonest. As for the philosopher’s comment that his student has beautiful handwriting, this blatantly flouts the maxims of quantity and/or relation—it is far too little information and is irrelevant. To make sense of the remark, and assuming that the writer is still playing the conversational game, we conclude that he is implying that his student is not a good philosopher.

But very frequently, adhering to the maxims gives rise to implicature too:

A: I am out of petrol.

B: There is a garage round the corner.

B here implicates that the garage round the corner *is open*, for he would be violating the maxim “be relevant” if it were not.

A: Smith doesn’t seem to have a girl friend these days.

B: He’s been paying a lot of visits to New York lately.

Again, B implicates that A has a girlfriend in New York: via the maxim of relevance. In both these cases, the speaker *implicates* that which he must be assumed to believe in order to preserve the assumption that he is observing the maxim of relation. But what is implicated may be false, yet the statement literally true.

There is much of interest and lasting value in Grice’s approach to these issues. There is also a change in philosophical orientation: Grice’s theory is concerned with language and meaning, but it is also a *theory*, of a quite general kind. For the early Oxford ordinary language philosophers—Ryle and Austin, for example—philosophical problems arise through the misuse of words, and are resolved or dissolved by clarifying the proper use of words. Philosophical investigations into meaning should be piecemeal, conducted on a case-by-case basis, and systematic theories of meaning are not required. By the mid-1960s—Grice’s William James Lectures in Harvard took place in 1967—it was clear that some at least of the ordinary language movement were becoming dissatisfied with the absence of systematic theorizing.⁸

12 Developments in Ethics

Barry Dainton

Meaning and metaphysics were by no means the only concerns of Oxford philosophers in the 1950s and 1960s, there was some very notable work done in ethics as well. In 1952 R. M. Hare published *The Language of Morals*, now recognized as a classic of analytic meta-ethics, and it was followed in 1963 by *Freedom and Reason*. On the first page of *Language* Hare makes his commitment to the linguistic approach clear: "Ethics, as I conceive it, is the logical study of the language of morals." It is not right and wrong per se that we are going to investigate, but certain features of the logic of moral discourse. Much the same approach is adopted in *Freedom and Reason* where Hare tells us:

Ethical theory, which determines the meanings and functions of the moral words, and thus the "rules" of the moral "game," provides only a clarification of the conceptual framework within which moral reasoning takes place; it is therefore, in the required sense, neutral as between different moral opinions. (1963, p. 89)

Hare followed Moore, Ayer, and Stevenson in rejecting naturalistic ethical theories: moral statements are not themselves factual in nature, nor are they deducible from solely factual premises. Hare followed Ayer and Stevenson (but not Moore) in rejecting nonnatural properties and the associated faculty of intuition. For him, evaluative terms fulfill a sui generis function in language, that of commending courses of action, and so cannot be analyzed in terms of words or expressions that do not themselves perform this function. In "Truth" Strawson had argued that the expression *is true* does not describe anything—it does not refer to a genuine property. If you say "S is true" you are not asserting anything over and above what you would have asserted simply by uttering "S," however, you are *doing* more: in using the truth predicate you are performing a specific speech act, that of affirming or endorsing that S. In a similar vein, according to Hare, moral expressions such as "good," "ought," and "right" do not refer to moral properties instantiated in the world, but are used to perform a distinctive sort of commendatory speech act.

Although in some respects Hare's position is similar to that of the emotivists, who held that the purpose of moral discourse was to impact on people's behavior by influencing their feelings, there are also important differences. On Hare's view, when we use moral terms we are *not* primarily trying to influence people's feelings or behavior. Instead, we are offering *guidance*. If I say "You ought to do X," I am telling you what to do, in response to an implicit (or explicit) question "What shall I do (in these circumstances)?" In response, you may in fact act in the way I have recommended, and indeed, I may well have intended this. But in understanding what is going on in such cases, it is crucial to distinguish what I have done, namely, recommend a course of action, from the *effects* or consequences of that speech act: your actually pursuing the course of action I have recommended. We saw earlier that Austin distinguished three aspects of speech acts: (a) the act *of* saying something, (b) what one does *in* saying it, and (c) what one does *by* saying it. In the current case, what I am doing *in* telling you that you should do X (i.e. that X is what I think you should do) is distinct from what I may accomplish *by* saying this (i.e. your doing X). If morality concerns action-guidance, in this sort of way, we have the beginnings of an answer to one of the main objections to emotivism. If moral discourse boils down to influencing behavior by the manipulation of feelings, it looks very much as though morality is ultimately nonrational. But if morality is all about recommending courses of action—offering guidance—then it is a different matter. If I tell you that you ought to do X, and you construe what I have said along these lines, then you can ask me for my *reasons* for recommending this course of action, and rational debate becomes possible.

Some of the logical positivists (e.g. Carnap at one stage) held that moral statements are disguised imperatives. My telling you that you ought to do X amounts to my saying "Do X!" Although Hare thinks this identification of moral claims with imperatives is mistaken, he does hold that they are intimately related. Imperatives and moral statements share the feature of both being *prescriptive*, and a typical moral statement entails an imperative, so if I say "You ought to do X," this carries with it the entailed imperative "Do X!" and if you accept that you ought to do X, then you are thereby *committed* to doing X. How do imperatives and moral judgments differ? If a few moments ago I indicated a preference for silence by telling you to "Shut up!" I am not thereby committing myself to telling you to stop talking on other occasions, even other occasions very similar to this one. Moral judgments, however, *do* involve a tacit commitment to consistency: they are "covertly universal,"

... which is the same as to say that they refer to, and express acceptance of, a standard which has an application to other similar instances . . .

When we commend an object, our judgment is not solely about that particular object, but is inescapably about objects like it. Thus if I say

that a certain motor-car is a good one, I am not merely saying something about a particular car. To say something about that particularly car, merely, would not be to commend. To commend . . . is to guide choices. Now for guiding a particular choice we have a linguistic instrument which is not that of commendation, namely the singular imperative. If I wish merely to tell someone to choose a particular car, with no thought of the kind of car to which it belongs, I can say "Take that one." If instead of this I say "That is a good one," I am saying something more. I am implying that if any motor car were just like that one, it would be a good one too; whereas by saying "Take that one" I do not imply that, if my hearer sees another car just like that one, he is to take it too. But further, the implication of the judgment "That is a good motor car" does not extend merely to motor-cars *exactly* like that one. If there were so the implication would be for practical purposes useless; for nothing is exactly like anything else. It extends to every motor-car that is like that one in the *relevant* particulars; and the relevant particulars are its virtues—those of its characteristics for which I was commending it, or which I was calling good about it. Whenever we commend, we have in mind something about the object commended which is the reason for our commendation. (1952, pp. 129–30)

And what goes for evaluations of motor cars extends to *moral* evaluations. The moral judgments that we make about (say) a particular action in a particular context must be grounded in certain features of that situation, and in making the judgments we commit ourselves to making the same judgment in all relevantly similar circumstances. If I say "You ought to do X" in a particular context, I am committed to prescribing the same course of action for anyone—myself included—who finds themselves in a relevantly similar context. Hence moral judgments have, for Hare, the distinctive feature of being *universalizable*. Not surprisingly, he often refers to his position as "universal prescriptivism."

Hare's noncognitivist approach may have been more sophisticated than any of its predecessors, but by the late 1950s noncognitivism itself was coming into question, and the questioning intensified with a cluster of hard-hitting papers published in 1958 by Philippa Foot and Elizabeth Anscombe. In "Moral Beliefs" and "Moral Arguments" Foot aimed to move moral philosophy in an Aristotelian direction. The noncognitivists held that there is no essential connection between facts and values, but from an Aristotelian perspective this is not the case: the good and the right are essentially bound up with human well-being and human flourishing, and virtuous actions are those that promote our well-being. Since there are ascertainable facts concerning human well-being and what promotes it, *some* facts do have moral consequences. In this passage

from “Moral Beliefs” Foot mocks those who deny that the good is entirely independent of our natures:

It is surely clear that moral virtues must be connected with human good and harm, and it is quite impossible to call anything you like good or harm. Consider for instance, the suggestion that a man might say he had been harmed because a bucket of water had been taken out of the sea . . . It would be just as odd if someone were supposed to say that harm had been done to him because the hairs of his head had been reduced to an even number. (1958–9, p. 94)

Far from its being up to us to choose the (moral) principles we live by, as Hare claims, the range of actions that can *conceivably* be regarded as right or virtuous—or wrong or harmful—is constrained by facts about our nature, or so Foot argues:

Is it even to be suggested that the harm done by a certain trait of character could be taken, by some extreme moral eccentric, to be just what made it a virtue? I suggest that such a man would not even be a moral eccentric . . . How exactly the concepts of harm, advantage, benefit, importance, etc., are related to the different moral concepts, such as rightness, obligation, goodness, duty and virtue, is something that needs the most patient investigation, but that they are so related seems undeniable, and it follows that a man cannot make his own personal decisions about the considerations which are to count as evidence in morals. (1958, pp. 510–11)

Anscombe’s “Modern Moral Philosophy” was also highly influential—and also stimulated interest in virtue ethics—but it was more iconoclastic. In her opening paragraph Anscombe tells us that she will be presenting three theses. The first is that “it is not profitable for us at present to do moral philosophy” and that we should put it aside until we have an adequate “philosophy of psychology” that at present we lack. Her second thesis is that “the concepts of obligation and duty—*moral* obligation and *moral* duty—and of what is *morally* right and wrong, and of the *moral* sense of ‘ought,’ ought to be jettisoned.” The final thesis is that there are no important differences between all “the well-known English writers on moral philosophy from Sidgwick to the present day” (1958, p. 1). In Anscombe’s eyes, the differences in doctrine between Sidgwick, Moore, Ross, Ayer, Stevenson, and Hare are minor compared to their flagrant and colossal errors.

As to why we should discard moral philosophy—or at least moral terms such as “right” and “ought”—it is Anscombe’s contention that these are

essentially legalistic in nature. For Aristotle a term such as “ought” did not have the modern sense of *being bound or obligated to do something*. The latter sense arrived only with Christianity and the notion of a divine lawgiver, which the Christians themselves took from the Torah. Hence these terms are entirely inappropriate within the secular framework that most contemporary moral philosophers adopt: it is “as if the ‘criminal’ were to remain when criminal law and criminal courts had been abolished and forgotten.” If we are to continue to do moral philosophy at all, we need to look back to Aristotle for a starting point:

In present-day philosophy an explanation is required how an unjust man is a bad man, or an unjust action a bad one; to give such an explanation belongs to ethics; but it cannot even begin until we are equipped with a sound philosophy of psychology. For the proof that an unjust man is a bad man would require a positive account of justice as a “virtue”. . . . The terms “should” or “ought” or “needs” relate to good and bad: e.g. machinery needs oil, or should or ought to be oiled, in that running without oil is bad for it, or it runs badly without oil. (1958, p. 4)

As for the essential similarity of recent English moral philosophers, Anscombe claims that they may have different conceptions of the nature of the good—Moore’s view is not shared by Stevenson or Hare—but they nonetheless all accept that the right action is the one that produces the best consequences. As a result, they are all committed to the view that it is impossible to rule out in advance that certain acts—for example, killing or torturing an innocent person—are always going to be wrong. In contrast, in the Hebrew-Christian ethical tradition there are moral absolutes: some acts are forbidden *in all circumstances*, irrespective of the consequences. Hence her conclusion: “if every academic philosopher since Sidgwick has written in such a way as to exclude this ethic, it would argue a certain provinciality of mind not to see this incompatibility as the most important fact about these philosophers, and the differences between them as trifling” (1958, p. 8). There are differences between Moore, Hare et al., but when viewed with anything resembling a proper historical perspective, these differences are altogether trivial.

Quite which approach to moral philosophy Anscombe is recommending in “Modern Moral Philosophy” is not entirely clear. As a practicing Roman Catholic, she is herself obviously an adherent of the older Hebrew-Christian tradition within which the distinctively moral senses of “ought” and “right” are fully intelligible, but she is well aware that a return to this tradition is not going to be an option for most contemporary ethicists.¹ Although she seems to believe a return to virtue ethics is the most promising alternative, in the final

paragraphs of the article she appears pessimistic as to the prospects of developing a viable moral system along such lines:

... it can be seen that philosophically there is a huge gap, at present unfillable as far as we are concerned, which needs to be filled by an account of human nature, human action, the type of characteristic a virtue is, and above all of human "flourishing." And it is this last concept that appears the most doubtful. (1958, p. 18)

An alternative approach, which Anscombe also considers, is grounding morality in a form of tacit social contract, as envisaged earlier by Hobbes, Locke, and Rousseau:

Just as we look at the law to find out what a man subject to it is required by it to do, so we look at a contract to find out what a man subject to it is required by it to do. Thinkers, admittedly remote from us, might have the idea of a *foedus rerum*, of the universe not as a legislator but as the embodiment of a contract. Then if you could find out what the contract was, you would learn your obligations under it. (1958, p. 14)

But if this approach offers something to those who would retain ordinary moral concepts, Anscombe foresees problems ahead for those seeking to develop such an approach:

... you cannot be in a contract without having contracted, i.e. given signs of entering upon the contract. Just possibly, it might be argued that the use of language which one makes in the ordinary conduct of life amounts in some sense to giving the signs of entering into various contracts. If anyone had this theory, we should want to see it worked out. I suspect that it would be largely formal; it might be possible to construct a system embodying the law (whose status might be compared to that of "laws" of logic): "what's sauce for the goose is sauce for the gander," but hardly descending to such particularities as the prohibition on murder or sodomy. Also, while it is clear that you can be subject to a law that you do not acknowledge and have not thought of as a law, it does not seem reasonable to say that you can enter upon a contract without knowing that you are doing so; such ignorance is usually held to be destructive of the nature of a contract. (1958, p. 14)

In any event, the challenges Anscombe lays down in her essay were taken up by others. Not only did virtue ethics start to flourish, interest in *all* areas of moral philosophy rapidly increased during the 1960s and 1970s, not least a

renewed interest in the contractual approach, culminating in Rawls' massively influential *A Theory of Justice* (1971).²

Like many of his predecessors in the contractualist tradition, Rawls begins with an imaginary scenario. He invites us to think of ourselves as free and equal people, tasked with agreeing on principles of political and economic justice to which we will all henceforth be committed. The parties to this "original position" are given a list of all the main doctrines pertaining to social justice from the main competing theories, and have to select which they want to apply to their own society. To ensure impartiality, Rawls imposes a "veil of ignorance": in selecting our principles we know nothing about our own personal characteristics, or social circumstances. We are to suppose ourselves ignorant of our sex, class, age, intelligence, state of physical well-being—everything that might conceivably be distinctive about us. Rawls goes on to argue that under these circumstances it would be rational to choose his two "principles of justice":

- (a) Each person has an equal claim to a fully adequate scheme of equal basic rights and liberties, which scheme is compatible with the same scheme for all; and in this scheme the equal political liberties, and only those liberties, are to be guaranteed their fair value.
- (b) Social and economic inequalities are to satisfy two conditions: (i) They are to be attached to positions and offices open to all under conditions of fair equality of opportunity; and (ii), they are to be to the greatest benefit of the least advantaged members of society.

The first principle is to have priority over the second, and in the latter (i) has priority over (ii). Rawls' *Theory of Justice*, together with Nozick's (rights-based and libertarian) *Anarchy, State and Utopia* (1974) would dominate debates in political philosophy for the next two decades.³ Scanlon's *What We Owe Each Other* (1998) is a more recent defense of contractualist ethics.

Hare is very unfashionable and Rawls is still a major figure—see Andre Moles' chapter on political obligation in this volume—but an important similarity between them is worth noting, namely the way that a liberal neutrality is built into both of their systems. In Hare it takes the form of universalizability, as the doctrine is found in *Freedom and Reason*. There it is not merely a matter of prescribing principles that contain only general terms, not names or particular references, as was the case in *Language of Morals*. It also involves putting yourself in the shoes of every person who might be affected by your action, and abstracting from your own preferences, desires, and metaphysical beliefs and choosing as if you were they. This is a very similar maneuver to Rawls' "initial position" behind the "veil of ignorance," where people do not even know their own conception of the good. In both cases, substantive values beyond neutrality seem to be in danger of being lost.

In their retrospective survey “Toward Fin de siècle Ethics: Some Trends,” Darwall et al. call this post-Harean period “The Great expansion”:

Slowly, the landscape of moral philosophy, which had become stark, even desiccated, during the final years of the reign of analytic metaethics, was being populated by a richer variety of views, many of which placed substantive and normative questions at the fore.

In the United States in particular, one such view became the reference point for all others, thanks in part to its systematic character and normative attractiveness: John Rawls’ *Theory of Justice*, with its method of “reflective equilibrium.” The narrowly language-oriented agenda of analytic metaethics was fully displaced. . . . A period that might be called “the Great Expansion” had begun in ethics.

In the Great Expansion a sense of liberation came to ethics. Moral philosophers shed the obsessions of analytic metaethics, and saw—or thought they saw—ways of exploring normative morality as a cognitive domain, without a bad philosophical conscience. The result was an unprecedented pouring of philosophical effort and personnel into ethics, which in turn spread out into the most diverse issues and applications. . . .

During the Great Expansion, moral intuitions (not Moorean insights into the Forms but substantive moral responses that strike us as compelling) flowed abundantly—occasionally urged on by a bit of pumping. Competing normative theories were “tested” dialectically against these intuitions in a procedure that appeared to be licensed by reflective equilibrium. (Darwall et al. 1992, pp. 122–3)⁴

It was not very long before new forms of cognitivism (or objectivism or moral realism) were being developed, for example, Sturgeon (1988), Railton (1989) Jackson and Pettit (1990), McDowell (1998), alongside new forms of noncognitivism—for example, Blackburn (1984) and Gibbard (1990).

One of the more significant aspects of the Great Expansion was a gradual move from the concern with defining “good” or “right” in their moral senses to investigating the wider issue of rationality in action in general. Ethics and moral philosophy are concerned with moral right and wrong, and hence moral reasons for acting (or not acting) in certain ways in certain circumstances. But in the empiricist tradition which can be traced back as far as Hume, and within which more recent non-cognitivists also work, there has long been a more general concern: whether anything could constitute a rational ground for action of any sort. An emotivist who thinks that moral values merely express subjective preferences, will probably think that acting on self-interest is entirely rational. He has every reason, he will assume, to avoid pain and seek happiness for himself, even if he has no such reasons for taking

steps to help others. But according to a sceptical line of thought, no facts can constitute rational grounds for action, not even facts about how pain feels, or what one's deepest desires or preferences are. The idea that reason can never by itself motivate actions derives from Hume's chapter "Of the influencing motives of the will," and it is also the message of Hare's "Pain and Evil" (1964) and "Wanting: some pitfalls" (1971).

The growing realization of the importance of this foundational issue has led to its occupying centre stage in recent debates. Ethics remains one of the most dynamic areas of analytic philosophy, and these issues of normativity and practical rationality are taken up by Ruth Chang in her chapter.

13 Davidson

Barry Dainton

Born in Springfield, Massachusetts, in 1917, dying in 2003, Donald Davidson was one of the most influential American philosophers of the postwar period. Early on he was interested in literature and the classics, and influenced by Whitehead—the latter had by then embarked on his later idealist phase. Davidson's work on classical philosophy was interrupted by war service, but he graduated from Harvard in 1949 with a dissertation on Plato's *Philebus*. At Harvard he met Quine and was powerfully influenced by him in turn. Henceforth he focused his energies on analytic philosophy. His early work in the 1950s was on experimental approaches to decision theory, and his work with Patrick Suppes led to the conclusion that it was not normally possible to pin down people's beliefs and preferences independently, with the consequence that there would always be a multiplicity of ways in which a given person's actions can be explained in terms of their reasons and preferences. From the 1960s onwards Davidson published a series of groundbreaking and influential essays in action theory, metaphysics (of events), philosophy of mind, and the philosophy of language. As gradually became clear, although seemingly self-contained, these papers contributed to the elaboration of a single systematic philosophical position.

Davidson's first major paper was "Actions, Reasons and Causes" (1963). Here he defends the view—now commonplace, then almost revolutionary—that the sorts of explanations we put forward when we explain people's actions in terms of their reasons, the explanations we are offering, are *causal* in nature. A key component in his argument for this conclusion was the thesis that events (such as actions) are particulars that can be described in more than one way. When described *as* an action, an event will not fall under any law-like regularity, but this does not rule out there being another description that applies to the event, and when the latter is described in *this* way it does fall under a law-like regularity, and consequently be considered as something that can both be caused (by other events), and be a cause (in its own right). Davidson can thus hold that explanations in terms of reasons will not usually involve strict law-like regularities, while also holding that there *are* such

regularities—at the level of events described in nonrational physical terms, say—underlying the rational explanations.

The more general thesis that the mental cannot be brought under strict laws, and so is irreducible to the physical, is a key component of his highly influential doctrine of the *anomalism* of the mental, which we will encounter again in Chapter 15, when we look in more detail at developments in the analytic philosophy of mind. We focus here on Davidson's views on meaning and translation, which are distinctive in their own way, and which had an almost explosive impact on the Oxford of the 1970s, where they displaced the until-then dominant ordinary language philosophy.

The *locus classicus* of the Davidsonian approach to language is the 1967 paper "Truth and Meaning."¹ Here Davidson argues that progress will be made in the philosophy of language if a proper *theory of meaning* can be developed. One might think that any philosophical account of language or meaning, of the kind (say) that Wittgenstein or Russell offer, could be considered to be "a theory of meaning." However, Davidson has something distinctive in mind: a systematic *axiomatic* theory of meaning. At the core of such a theory there will be a finite number of axioms that, for some particular language, will specify the semantic properties of words and their modes of combination. These axioms yield *theorems*, and the latter will tell us the semantic properties of all the sentences that can be expressed in a particular language. The prime advantage of a theory of meaning taking this axiomatic form is that it is *compositional*, that is, the meanings of complete sentences is wholly determined by the meanings of their component words and their manner of composition.

As for why Davidson (and other philosophers of language) think compositionality important, it suffices to reflect on two facts that any account of natural language has to take into account. First of all, typical speakers can produce and understand sentences they have never heard before—"Lenin was an aardvark's uncle," for example. We can all dream up entirely novel sentences—some silly, many not—effortlessly. Secondly, as infants we rapidly master languages that are entirely new to us, and we do so after hearing only a minuscule fragment of the total number of sentences that our mother tongues actually contain. Our ability to learn languages in the way that we do, and comprehend (and produce) entirely novel sentences would be miraculous—if not wholly impossible—unless the meaning of sentences were determined in a rule-governed way by their mode of composition and the semantic properties of their component words. Moreover, the rules that allow us to work out the meanings of whole sentences from their constituent parts must clearly be finite in number: if learning a language required the mastery of an infinite number of rules none of us would ever manage it. Hence the importance of compositionality and recursiveness.

So far so good. We now know roughly what an axiomatic theory of meaning involves, but another question now arises: what sort of *theorems* will it generate? Let us suppose that the target (or *object*) language is *L*, and our theory is framed in a second language, our *metalanguage*. One possibility, a very natural one, on the face of it, is:

s means that **p**

Where **s** is a name that refers to a single sentence in our object language, and **p** a sentence in our metalanguage that states the meaning of **s**. If French is the object language, and English our meta-language, then a typical theorem of a type of this sort would be:

“La neige est blanche” means snow is white.

But Davidson rejects this idea. Like Quine, he is suspicious of intensional concepts such as meaning, and in any case, since theorems of this form would themselves presuppose the concept *meaning* they would not provide as much insight into the nature of this concept as theorems that do not presuppose it. So we need an alternative. Theorems of this form are one option:

s if and only if **p**

But as Davidson notes, **s** is not itself a *sentence*, it is a letter functioning as the name of a sentence. Since the biconditional “if and only if” links whole sentences, we will have to add a predicate to **s** to create a genuine sentence on the left-hand side. If we do this we get this:

s is *X* if and only if **p**

Now, since **s** names either **p** itself, or a translation of **p**, what must the predicate *X* be such that “snow is white” has the property *X* if and only if snow is white? The answer is not difficult to locate: *truth*. If *X* is interpreted as “is true,” then if **s** names a true sentence, then **p** will be true, and “**s** is true” will be a true sentence. If **s** names a false sentence, then **p** will be false, and “**s** is true” will be false. More generally, the left- and right-hand sides of *all* biconditionals of this form have the same truth value if (and only if) we interpret the predicate *X* as “is true.” Hence we arrive at the final form of the desired theorems:

s is *T* if and only if **p**

where “*T*” is the truth-predicate. Sentences of this form are called *T-sentences*, and they are wholly extensional. *T-sentences* effectively give us the *truth-*

conditions of sentences in the object language. They tell us the conditions under which object sentences are true.

Davidson concedes that truth-conditions of the sort given by individual T-sentences are not identical in meanings, since *all* that is required for the truth of a T-sentence is that the linked sentences have the same truth value, and this is not much. What he does claim is that if a T-sentence is part of an entire (true) truth-theory for a whole language, and this theory fits the empirical evidence as well as any theory can, then we are justified in thinking that the T-sentences this theory yields will provide us with adequate *interpretations* of the target sentences in the unfamiliar language. But more on this shortly.

Davidson now has his theorems—or at least their general form. The next issue to settle is the character of the axioms of our theory of meaning. In an extensional language, the truth of sentences is dependent upon only the extensions of the component expressions. Since Davidson insists on extensionality, our axioms will assign extensions to the relevant expressions: (a) for singular terms they will specify referents, for example, Shakespeare to “Shakespeare,” snow to “snow”; (b) for predicates they will specify what the predicates apply to: “is white” applies to all and only white things; “is taller than” applies to all and only ordered pairs of objects where the first member of the pair is taller than the second.

At this point an appealing shortcut becomes possible. Alfred Tarski’s most celebrated achievement was his a compositional *theory of truth* for formal languages—as found in the seminal “The Concept of Truth in Formalised Language,” published in 1933. His theory “defines” truth for a language by virtue of possessing axioms that generate true T-sentences of the form “*s* is T if and only if *p*.” Tarski’s idea, roughly, was that for a given language *L*, each of these T-sentences in *L* constitutes a *partial* definition of “true sentence” for *L*, and hence that a theory that generates *all* the T-sentences for *L* will be a complete—or “materially adequate”—definition of truth for that language.² As well as furnishing us with theorems of the right form, Tarski’s theory is also fully extensional: its axioms assign only extensions to the terms in *L*. Davidson concludes that we need look no further for our theory of meaning: *if* an adequate theory of meaning will produce T-sentences, then we already have the basic logical form of the theory we are looking for in Tarski’s truth-theory.

When Tarski developed his truth-theory, he took the concept of meaning for granted. His theory defines truth by virtue of having axioms that generate true T-sentences of the form “*s* is T if and only if *p*.” Since in Tarski’s theory the sentence *p* is either *s* itself, or a translation of *s*, we can be sure that his T-sentences are true. What Davidson is proposing is altogether different. Whereas Tarski could take for granted the notion of meaning and sameness of meaning, Davidson, in attempting to develop a *theory* of meaning, certainly

cannot. What he does, instead, is take the notion of *truth* for granted, as primitive, and uses the T-sentences as a way of approaching meaning. In effect, his approach is the opposite of Tarski's:

In Tarski's work, T-sentences are taken to be true because the right branch of the biconditional is assumed to be a translation of the sentence truth-conditions for which are being given. But we cannot assume in advance that correct translation can be recognized without pre-empting the point of radical interpretation. What I propose is to reverse the direction of explanation: assuming translation, Tarski was able to define truth; the present idea is to take truth as basic and to extract an account of translation or interpretation. The advantages, from the point of view of radical interpretation, are obvious. Truth is a single property which attaches or fails to attach, to utterances, while each utterance has its own interpretation; and truth is more apt to connect with fairly simple attitudes of speakers. (1984, p. 134)

Davidson's approach is bold and ingenious, but as he is well aware, there is a considerable obstacle to be circumvented or overcome. As was noted above, it does not take much to make Davidson's theorems true. Sentences of the form "*s* is T if and only if *p*" are biconditionals. Any biconditional sentence is true if and only if the two sentences on either side of it have the same truth value. But since this is *all* that is required for a T-sentence to be true, it follows that a T-sentence such as:

"Snow is white" is T if and only if grass is green.

is true, since both left- and right-hand sentences have the same truth value: true. Yet, it can hardly be said that this T-sentence gives the *meaning* of "snow is white," quite the reverse. The main reason Davidson gives for preferring an extensional theory of meaning is that such a theory is adequate: we do not *need* to appeal to intensional notions like meaning in order to develop an adequate theory of meaning. Yet, on the face of it, extensional T-sentences will not give us what we need. A truth theory is true for a language L if it provides a true T-sentence for every sentence of L. Given that so little is required for a T-sentence to be true, there will be many true T-theories for any language L, but very few of these will provide good interpretations of the sentences of L.

Clearly, what Davidson must do if his approach is to be remotely adequate is place additional constraints, over and above extensional adequacy, on what counts as an adequate T-theory for the purposes of a theory of *meaning*, that is, a T-theory whose theorems we have good reason to suppose provides us

with good interpretations or translations of sentences in *L*. Some of his most distinctive doctrines emerge in his attempts to do precisely this.

Many erroneous translations will be eliminated thanks to the systematic nature of *T*-theories. Since a theory that entails that “snow is white” is true iff grass is green would also entail that “snow melts in the sun” is true iff grass melts in the sun, flaws such as these could be detected. However, there are problematic translations that structural considerations do not rule out. Davidson’s solution is to embed his truth-theory into a full-scale theory of interpretation; the required additional constraints on *T*-theories will emerge from this, or so Davidson claims.

We are once again invited to imagine ourselves confronted with the task of developing a theory of meaning for an unknown language; the envisaged task is similar to Quine’s project of radical translation, even if Davidson prefers to talk of “radical interpretation” (the projects *are* different, as we shall see). When engaging in radical interpretation we observe subjects using sentences in various situations—whether in the jungle or in the home—and on this basis we develop hypotheses about what their utterances mean. Recall Quine’s example: we see a rabbit running past, a native-speaker pointing to it shouting: “Gavagai!” and we write down, as a tentative translation “Rabbit!” There are, however, some significant differences between Quine’s conception of radical translation and Davidson’s radical interpretation. First, Davidson has no truck with relating patterns of sentence-usage to patterns of sensory stimulation (surface irritations): he relates usage directly with objects and events in the world. Second, in Davidsonian radical interpretation, the evidence goes into constructing a truth-theory for the target language: this theory provides the basic (Tarskian) structure of the interpretation. Last but not least, Davidson holds that interpreters should make full use of the “Principle of Charity” roughly, we should assume that the speakers of our target language are as rational as we are. This may seem innocuous enough, but in Davidson’s hands it has far-reaching consequences.

Quine himself advocated employing the Principle of Charity in his own account of radical translation. If a proposed translation of the words in the target language for logical connectives leaves its speakers looking wildly irrational, the correct course—suggests Quine—is to assume that our translation has gone astray, or as he puts it “one’s interlocutor’s silliness, beyond a certain point, is less likely than a bad translation” (1960, p. 59). But whereas Quine restricts the domain to which the Principle of Charity should be applied to the basic canons of logic, Davidson applies it “across the board”: to all utterances, on all topics. He holds that when we undertake radical interpretation, we must assume that (a) the speakers of the target language *L* share our canons of rationality, and (b) that they share many of our basic *beliefs*. For

Davidson this is not just a piece of methodological advice or a useful heuristic, but a necessary condition for counting *L*-speakers as speakers of any kind of language, and indeed, as possessing of psychological states such as beliefs and intentions.

For Davidson rationality and belief are closely connected. We regard it as rational to believe what is true, and we regard *our own* beliefs to be true and rationally grounded—at least our beliefs about our immediate surroundings, the shape of the Earth, the relationship between clouds and rain, and so forth. Consequently, in making sense of the speakers of a language we are trying to understand and interpret, we are bound to ascribe to these speakers as many of *our own* basic beliefs—or at least, the beliefs we would have if we were in the natives' shoes—since these are the only beliefs we will find it rational to suppose the natives (and anyone else) could have, in their circumstances. Thus it is that maximal agreement is a presupposition of the interpretive project: the interpretation that renders others peoples' overall behavior patterns *maximally* intelligible will ascribe to them the maximal number of beliefs that, by our lights, it is rational for them to hold in the relevant circumstances.

There is a Kantian flavor to this: in effect, Davidson is putting forward a *transcendental argument*, one that supplies necessary conditions for the possibility of successful interpretation, the upshot of which is that anyone we manage to understand—anyone whose utterances we can reasonably attribute meaning to—will turn out to be pretty similar to us in terms of their standards of reasonableness and basic beliefs. Here is Davidson himself on these points:

... the aim is not the absurd one of making disagreement and error disappear. The point is rather that widespread agreement is the only possible background against which disputes and mistakes can be interpreted. Making sense of the utterances and behaviour of others, even their most aberrant behaviour, requires us to find a great deal of reason and truth in them. To see too much unreason on the part of others is simply to undermine our ability to understand what it is they are so unreasonable about. If the vast amount of agreement on plain matters that is assumed in communication escapes notice, it's because the shared truths are too many and too dull to bear mentioning. What we want to talk about is what's new, surprising, or disputed. (1984, p. 153)

And again:

What makes interpretation possible ... is the fact that we can dismiss *a priori* the chance of massive error. A theory of interpretation cannot be correct that makes a man assent to very many false sentences; it must

generally be the case that a sentence is true when a speaker holds it to be. . . . No simple theory can put a speaker and interpreter in perfect agreement . . . The basic methodological precept, is, therefore, that a good theory of interpretation *maximizes* agreement. (1984, pp. 168–9)

In “Mental Events” he is more lyrical: we must “try for a theory that finds him consistent, a believer of truths, and a lover of the good (all by our own lights, it goes without saying)” (1970, p. 222). If we suppose that Davidson is right about this, there are a number of consequences.

First of all, the across-the-board application of the Principle of Charity leads, or so argues Davidson, to a great reduction in the extent to which interpretation is indeterminate. Because Davidsonian interpretations maximize agreement, many translations that would be acceptable to Quine are unacceptable for Davidson. We also have a solution to the “biconditional problem.” Since the theorems of Davidson’s meaning-theory are of the form “*s* is true if and only if *p*,” and all it takes for sentences of this form to be true is for the sentences *s* and *p* to have the same truth-value, there are many true truth-theories that are very poor interpretations (e.g. which interpret “*la neige est blanche*” as meaning that grass is green). These surplus-to-requirement theories will be eliminated—or greatly reduced in number, at least—if we constrain our interpretations by the Principle of Charity in the way Davidson recommends.

There are also important implications for crosscultural heterogeneity. According to one familiar form of cultural relativism it is possible that different cultures (or speech communities) might vary greatly in their “conceptual schemes.” That is, the basic concepts that they use to make sense of their experience, and their basic beliefs, might be very different from ours. In his essay “The Very Idea of a Conceptual Scheme” (1974) Davidson argues that it is incoherent to suppose conceptual schemes might vary in this manner. How could we ever discover if a particular community was working within a conceptual scheme drastically different from our own? We would first have to understand their language, so as to know how they viewed things, and discover which concepts they employed. But: if interpretation necessarily involves charity, as Davidson maintains, any interpretation we reach of their language will be a rationalization of their behavior, and this rationalization will involve our ascribing to the target community principles of reasoning, and basic beliefs that we find reasonable, at least in the main. If we succeed in interpreting their language, then the natives will emerge as largely reasonable by our standards, that is, as working within a conceptual scheme that is broadly similar to our own. Hence given Davidson’s approach, the very idea of drastically different conceptual schemes is simply *incoherent*. It is not that it would be difficult to arrive at an understanding of an alien scheme, the very

idea of such a scheme makes no sense. Anyone who speaks a language at all, who operates with any concepts at all, operates with concepts pretty similar to our own.

Thirdly, but again not least, there are epistemological implications. Radical or global skepticism is the doctrine that we are not justified in supposing that any or most of our beliefs are true. For all we know, *all* our beliefs could be false. This is the kind of skepticism introduced by Descartes, and in the centuries since then a vast amount of effort has gone into trying to refute it: by showing that we are justified in believing that at least some of our beliefs are true. But this has proved to be a difficult task. According to Davidson, global skepticism of this sort is simply incoherent, and so refuted. This conclusion follows directly from the Principle of Charity: the existence of meaningful content in sentences or mental states necessarily requires the bulk of our simple beliefs to be true. Provided that we speak a language, think thoughts, and possess beliefs, we can be assured that our beliefs are mostly true—some may be false, but most cannot be. If our beliefs *were* mostly false, Davidson argues, then none of utterances or propositional attitudes could possess intentional content. For in this case it would be impossible for even an omniscient interpreter to interpret them. But of course our utterances and propositional attitudes do possess content; we interpret others successfully, and others interpret us successfully. Consequently, we can be confident that our own and other's beliefs about the world are mostly true.

These are remarkable results, and needless to say they have provoked a considerable amount of discussion.³ Let us take a step back from the fray. One thing at least is very clear: with Davidson's work we have yet another example of an analytic philosopher deriving potentially powerful results from reflections on language and logic. So far, so familiar. But there are traces of something new here too, signs of an alteration in fundamental orientation.

Davidson's more general conclusions are certainly interesting, but they can easily seem almost too strong—too interesting. The claim that there is no *possibility* of our basic beliefs failing to correspond with reality is such a strong claim that it can easily seem absurd. It might seem all the more absurd given that this conclusion derives from reflections on the nature of meaning. Language, as we all know, is fully capable of expressing false as well as true propositions. How, then, could the mere existence of a language such as ours guarantee that the propositions that express our basic beliefs are mostly true? Much of the intuitive resistance to Davidson's position largely probably stems from a pre-theoretical attachment to strong metaphysical realism: the idea that there is a reality that is independent of our beliefs, and one that could quite conceivably be very different from how our beliefs represent it. Davidson himself claims that his position is neither realist nor antirealist, as these terms are usually construed.⁴ However, given that he holds that truth simply is what,

on the whole, we believe—that reality itself cannot fail to conform to how we conceive and believe it to be—a strong case can be made for taking his position to a form of antirealism.⁵ Indeed, if the real is constrained by the limits of *our* concepts and beliefs in the way Davidson maintains, a case could be made for saying that Davidson renders reality mind-dependent in a way that is not dissimilar to some idealists. In which case, by the 1980s, at least one current in analytic philosophy was moving in a direction that would, no doubt, have alarmed the Russell and Moore of 1900.

14

Kripke and Putnam

Barry Dainton

In the 1970s there were a number of noteworthy new developments in the philosophy of language, involving a number of soon-to-be-prominent philosophers, but we will confine ourselves here to taking a very brief look at (some) of the work of just two: Saul Kripke (1940–) and Hilary Putnam (1926–), both currently emeritus professors at Princeton and Harvard respectively. These contributions are noteworthy both for their intrinsic interest, and their wider impact on other areas of the subject, such as the philosophy of mind and metaphysics.¹

One notion in particular played a key role in these developments in the 1970s, namely the idea of *direct reference* and the significance of extending such reference to the external world.² The plot ran as follows. Frege and Russell both subscribed to what could be viewed as a “descriptivist” theory of the meaning of proper names: both rejected the naïve view that the meaning of a proper name *just is* the object it refers to; both held that names have a descriptive sense or meaning, and that such expressions refer by virtue of this descriptive sense: they pick out the object that satisfies the associated descriptions. Frege and Russell also agreed that proper names—at least of the ordinary language variety—can contribute to contentful sentences even when they fail to refer to anything. But Kripke argued—for reasons we shall outline below—that there could not be an adequate descriptive theory of the functioning of names. Reference to objects is not mediated by a description or a Fregean sense, but is unmediated or direct; and the only *mode of presentation* of the object is the object itself. Putnam extended this claim to apply not only to proper names but also to natural kind terms, so there could not be any descriptive definition of terms such as “tiger,” “water,” “gold,” etc., in terms of the criteria by which we recognize such things: the meaning of these terms is determined by the real essences of such things, as revealed by science. At first this was a theory in semantics, but it took little imagination to see that it naturally extended to the thoughts that we think using these terms. So there was a move from what one might regard as the relatively harmless idea word that meaning can depend on something beyond the knowledge of the speaker, to the more dramatic claim that a subject’s thoughts are

conditioned by factors outside the mind or consciousness of the subject. This is the dramatic theory of *psychological externalism*. This story unfolds in the following way.

In a series of lectures in Princeton in 1970, Saul Kripke launched an attack on several (then) current orthodoxies, descriptivism included; the material in the lectures would eventually be published as *Naming and Necessity* (1980).³ The force of Kripke's main lines of argument can easily be appreciated. Let us take "Shakespeare" to be our example of an ordinary language proper name. If we assume that some form of descriptivism is true, the meaning of "Shakespeare," for a given speaker, will amount to a cluster of descriptions, for example, "famous English playwright of the late 16th to early 17th centuries, the author of *Macbeth*, *Hamlet*, etc., who had a partnership in the Globe Theatre." Now consider these claims:

S1 Shakespeare, if he exists, wrote *Romeo and Juliet*, *Hamlet*, and *Macbeth*.
If anyone is an English playwright who is sole author of *Hamlet*, *Macbeth*, and *Romeo and Juliet*, then he is Shakespeare.

If the descriptivist theory were true, we could replace each occurrence of "Shakespeare" with the descriptive equivalent, yielding the following:

S2 The author of *Hamlet*, *Macbeth*, and *Romeo and Juliet*, if he exists, wrote *Hamlet*, *Macbeth*, and *Romeo and Juliet*. If anyone is the sole author of *Hamlet*, *Macbeth*, and *Romeo and Juliet*, then this person is the sole author of *Hamlet*, *Macbeth*, and *Romeo and Juliet*.

Descriptivism leads to the conclusion that the S1 sentences are in fact analytic truths, that is, true by virtue of the words they contain, in which case they could not possibly be false. Yet as Kripke points out, this seems wrong. Surely it is possible that Shakespeare (the man) could have lived, but become a lawyer instead of a playwright. If so, the first sentence in S1 is false. Moreover, it is possible that someone else wrote the three plays mentioned—perhaps Francis Bacon was their author—in which case the second sentence in S1 would also be false. There are possible worlds where Shakespeare did not write any plays, and there are possible worlds where the plays were written, but by someone else—or so it seems natural to suppose—but the descriptivist theory is incompatible with these modal facts. The underlying point that Kripke is bringing out here is that we can use a name like "Shakespeare" to refer to one and the same individual in various possible but nonactual situations. In these counterfactual scenarios Shakespeare does not have the various distinguishing features he has in this world, yet we can still succeed in referring to him—indeed, we can

do so effortlessly. We would not be able to do this if “Shakespeare” simply meant “the individual with properties *x*, *y*, *z*.”

Kripke went on to make an epistemological point. If the descriptivist theory is true, then we should be able to know the truth of S1 just by reflecting on the concepts involved: it should be part of the concept “Shakespeare” that the person referred to wrote Hamlet, Macbeth, etc. But surely this too is wrong. In actual fact, we cannot come to know that he wrote these plays simply by analyzing the concept “Shakespeare”—we need to do some empirical research. Moreover, it could turn out that Shakespeare *did not* write these plays; evidence for this could emerge. But this would be inconceivable *if* Shakespeare simply meant “The author of Hamlet, Macbeth, etc.”

In a more direct vein, Kripke also argues that the descriptivist theory simply gets the semantic facts wrong: there are cases in which the reference it attributes to a name is the wrong reference. Suppose you are someone who knows nothing about Gödel save that he was the logician who discovered the incompleteness of arithmetic. Let us further suppose that Gödel was not in fact the discoverer of the theorem named after him, but instead

[a] man named “Schmidt,” whose body was found in Vienna under mysterious circumstances many years ago, actually did the work in question. His friend Gödel somehow got hold of the manuscript and it was thereafter attributed to Gödel. On the view in question, then, when our ordinary man uses the name “Gödel,” he really means to refer to Schmidt, because Schmidt is the unique person satisfying the description “the man who discovered the incompleteness of arithmetic”. . . . So, since the man who discovered the incompleteness of arithmetic is in fact Schmidt, we, when we talk about “Gödel,” are in fact always referring to Schmidt. But it seems to me that we are not. We simply are not. (1980, p. 84)

If your intuitions are the same as Kripke—if you can make sense of the scenario he describes *as* he describes it—then it is difficult to see how the names in question, at least as you use them, could function in the way the descriptivist maintains.

That Kripke’s criticisms pose a serious challenge for descriptivism is not in dispute; whether any form of descriptivism can successfully meet these challenges continues to be much debated—opinions on this are divided.⁴ Those who believe the descriptive approach should be rejected find themselves confronted with problems, some old, some new. What is it that connects a name (when used on a particular occasion) with the particular object to which it refers? If it is not the satisfaction of one or more definite descriptions

associated with the name, we need an alternative account. Kripke suggested that the answer lies in causal-historical factors:

Someone, let's say a baby, is born; his parents call him by a certain name. They talk about him to their friends. Other people meet him. Through various sorts of talk the name is spread from link to link as if by a chain. A speaker who is on the far end of the chain, who has heard about, say, Richard Feynman, in the market place or elsewhere, may be referring to Richard Feynman even though he can't remember from whom he first heard of Feynman or from whom he ever heard of Feynman. He knows that Feynman was a famous physicist. A certain passage of communication reaching ultimately to the man himself does reach the speaker. He is then referring to Feynman, even though he can't identify him uniquely. (1980, p. 91)

The picture Kripke sketches is simple and appealing, but it needs further refinement to deal with various problem-cases; see Evans 1973, 1982. This issue aside, if we follow Kripke and deny that names are semantically equivalent to descriptions, we are faced with the various puzzles that troubled Frege and Russell. What sort of meaning (if any) do bearerless names possess? How can identity statements such as "Clark Kent is Superman" be informative? Why can coreferring names not be substituted for one another in belief-reports? Those who favor some form of descriptivism have well-established solutions to these problems; those who reject descriptivism have to devise new solutions, and although a good deal of innovative work has been expended on these problems since the 1970s, these issues remain very much alive.⁵

Another notable Kripkean innovation is the distinction between two types of referring expressions (or "designators"), which he introduced thus: "Let's call something a *rigid designator* if in every possible world it designates the same object, a *nonrigid* or *accidental designator* if that is not the case" (1980, p. 48) It is clear that definite descriptions are nonrigid, for example "The US President in 1970" actually denotes Nixon, but it could have denoted someone else, Humphrey, for instance, if he had won the presidential election instead, and Nixon had gone on to become a talk show host instead of gaining the presidency. But proper names are different: they are rigid designators, or so Kripke argues. When we use a name, such as "Nixon" to make counterfactual claims—claims about what might have been the case, but is not—we are typically referring to the same individual as we are referring to in the actual world. So if one says "Nixon might not have been called 'Nixon,'" what we are claiming is something like this: "In some possible world, Nixon (the individual who is so-called in this world) was given a different name by his parents."

Of course, this divergence in the mode of functioning of proper names and definite descriptions could not arise if the descriptivist view of the former were true, but Kripke has already argued against the descriptivist approach, as we have just seen.

Kripke goes on to put the distinction to work, and introduces a number of identity statements featuring coreferential rigid designators: "Hesperus is Phosphorus," "Superman is Clark Kent," "gold is the element with atomic number 79," "light is a stream of photons," "water is H_2O ," and "pain is C-fibre activity." Prior to Kripke it was generally assumed that identity statements such as these were contingent, that is, that they could have been false, that they *are* false in other possible worlds. At first view, this can seem perfectly plausible: is it not easy to imagine a world where Clark Kent has ordinary powers and Superman is someone other than Clark Kent? Or a world where gold is not made up of the same sorts of atoms, but looks and behaves in just the same way? But if the relevant referring expressions are rigid designators these scenarios are not really possible: "Superman" and "Clark Kent" refer to the same object in this world, if they are rigid they must refer to the same object in every possible world. In which case there is no possible world where Superman is anyone other than Clark Kent. This is not to say that there are not worlds where there is someone very similar to Superman in appearances and powers but is not Clark Kent. There are many such worlds. The claim is simply that these people are not the Superman we know. The same holds for the other pairings Kripke mentions: "heat" and "motion of molecules," "gold" and "element with atomic number 79," and so forth.

This way of construing identity statements has broader consequences. First of all, it provides a new rationale for essentialism. According to essentialists, the properties of things divide into two categories: those that they possess essentially and so cannot exist without, and those that they possess only contingently and so can exist without. This metaphysical doctrine had fallen out of favor, largely due to the influence of Quine.⁶ But Kripke supplies this venerable doctrine a new lease of life. If the statement that "gold is the element with atomic number 79" is necessarily true, it is true in all possible worlds, and hence there is no world where gold is anything other than the element whose atomic number is 79. It looks very much as though this property is an essential property of gold—whereas many of its other properties, such as being so-and-so's favorite metal, are merely contingent. Hence if Kripke's analysis is correct, science is not just discovering contingent truths, it is discerning the essential natures of things. Moreover, it is philosophy that is revealing *what it is* that science is really doing when it uncovers the underlying physical nature of things.

Second, in the period when Kripke was writing it was generally accepted that necessary truths were analytic, and so knowable a priori. On Kripke's

analysis, although statements such as “lightning is an electrical discharge” and “water is H_2O ” are necessary truths, it is not remotely plausible to hold that they are analytic and a priori. Statements such as these are *theoretical* identifications: we cannot discover their truth merely by reflecting on the meanings of the relevant terms; rather, they are substantive (and often surprising) scientific discoveries. As such they are clearly *non*-analytic and a posteriori. So Kripke is providing us with a previously unrecognized—and obviously important—class of truths: those that are a posteriori and necessary. Of the several intriguing consequences of this, not the least is the threat to those who hold that it is the proper (and distinctive) role of philosophy to (a) discover truths that are necessary rather than merely contingent, and (b) do so by armchair conceptual analysis. If Kripke’s views on the necessary a posteriori are correct, in a broad range of cases it is scientists—not philosophers—who are the ones revealing metaphysically necessary truths, and these truths are not discoverable by a priori conceptual analysis. Those analytic philosophers who had found themselves able to resist Quine’s attack on the a priori now had a new challenge to meet.

Over the course of a long career—much of it spent at Harvard—Hilary Putnam has done important work in many areas of the subject: logic and the philosophy of mathematics, the philosophy of science, epistemology, ethics, metaphysics, and the philosophy of mind. As well as having a broad range of interests, Putnam’s views on many central issues have evolved—for example, the robust scientific realism he defended early on gave way to an “internal realism” according to which there is no one best way to characterize reality, but he later rejected this in favor of a Pragmatist-inspired direct realism. We will be ignoring all of this here to focus on just one line of argument in just one of his papers, “The Meaning of ‘Meaning’” (1975).

Putnam begins by isolating two key assumptions of the “myth-eaten” traditional theories of meaning that he will be criticizing: (a) that knowing the meaning of a word is just a matter of being in a certain psychological state, (b) that the meaning of a word (its “intension”) determines its extension (or reference), so that sameness of intension entails sameness of extension. Expanding on this in introduction to *The Twin Earth Chronicles* (Pessin and Goldberg 1996) some 20 years later,⁷ he tells us it had hitherto been assumed by almost all philosophers

that the idea in the mind, or possession or recollection of the idea by the mind determines the extension of the “name” associated with the idea or concept: a name, say “dog,” is *true of* a particular thing inasmuch as that particular thing falls under the concept in the mind, or the concept recollected by the mind. In short, it is a feature of all these views that *one individual in isolation* can, in principle, grasp any concept whatsoever,

and that individual's grasp of his or her concepts totally determines the extension of all the individual's terms. *Knowledge of meanings is private mental property*. (p. xvi)

Putnam acknowledges that Frege and Carnap (not to mention Russell and Moore) vehemently rejected the notion that concepts are *mental* in nature: rather, they are abstract entities that are publically available. Even so, the *grasping* of these abstract entities is a psychological act, which means—even for Fregeans—that understanding the meaning of a word is a matter of being in a particular kind of psychological state.

In arguing against this “traditional” view of meaning Putnam advances a number of considerations, among which is a now-famous thought-experiment:

That psychological state does not determine extension will now be shown with the aid of a little science fiction . . . we shall suppose that somewhere in the Galaxy there is a planet we shall call Twin Earth. Twin Earth is very much like Earth; in fact, people on Twin Earth even speak *English*. In fact, apart from the differences we shall specify, the reader may suppose that Twin Earth is exactly like Earth. . . .

One of the peculiarities of Twin Earth is that the liquid called “water” is not H₂O but a different liquid whose chemical formula is very long and complicated. I shall abbreviate this chemical formula simply as XYZ. I shall suppose that XYZ is indistinguishable from water at normal temperatures and pressures. In particular, it tastes like water and it quenches thirst like water. Also, I shall suppose that the oceans and lakes and seas of Twin Earth contain XYZ and not water, that it rains XYZ on Twin Earth and not water, etc.

If a spaceship from Earth ever visits Twin Earth, then the supposition at first will be that “water” has the same meaning on Earth and on Twin Earth. This supposition will be corrected when it is discovered that “water” on Twin Earth is XYZ, and the Earthian spaceship will report somewhat as follows: “On Twin Earth, the word ‘water’ means XYZ.” (1975, p. 223)

Putnam goes on to suggest that a Twin-Earther visiting Earth would behave in a similar (symmetrical) way: initially they would assume that “water” had the same meaning as it has on their home planet—namely XYZ—but they would later report back that it actually means H₂O. Putnam now asks us to roll back time, to around 1750, that is, to a period when chemistry was not very developed on either Earth or Twin Earth, and the inhabitants of both planets were entirely ignorant of the fact that water on Earth consists of hydrogen

and oxygen, and of XYZ on Twin Earth. He envisages two typical speakers, an Earthling Oscar₁, and a Twin-Earther Oscar₂, who are exact duplicates in physical appearance, feelings, thoughts, interior soliloquy, and so forth, and who act in precisely the same sorts of ways:

Oscar₁ and Oscar₂ understood the term “water” differently in 1750 *although they were in the same psychological state*, and although, given the state of science at the time, it would have taken their scientific communities about fifty years to discover that they understood “water” differently. Thus the extension of the term “water” (and in fact, its “meaning” in the intuitive preanalytical usage of that term) is *not* a function of the psychological state of the speaker by itself. (ibid., p. 224)

Putnam’s reasoning is, on the face of it at least, difficult to resist. It does seem plausible, intuitively, that Oscar₁ and Oscar₂ mean something different by “water.” If they are also indistinguishable psychologically, then since they mean—or at least refer—to different things when they use the term “water,” it seems meaning (or at least reference) is not determined by their psychological states alone, but by a combination of these and external (extra-mental) environmental factors. It is because we are on Earth rather than Twin Earth that we refer to H₂O rather than XYZ when we talk about “water.” Hence Putnam’s conclusion: “Cut the pie any way you like, ‘meanings’ just ain’t in the head!” (1975, p. 227).⁸

However, the issue of whether Earthers and Twin-Earther’s *are* psychologically indistinguishable is not entirely straightforward. Earlier in “The Meaning of ‘Meaning’” Putnam draws a distinction between “narrow” and “broad” mental states. Consider these two scenarios: (a) you are a normally embodied person, and things are as they seem, (b) you are an envatted brain, and your perceptual experiences are caused by stimulations to the nerve endings extruding from your brain. Since in both (a) and (b) everything is the same phenomenologically, is it not clear that that in *some* sense the mental states of the individual (you) is exactly the same in both scenarios? Intuitively it is, and Putnam calls mental states that are independent of the state—or even the existence—of the wider world *narrow states*. However, he points out that there are some mental states that are not world-independent, and calls these *wide states*. For example, if I am jealous of Sam, then Sam exists; “being jealous of” is a relational property that links me to a particular person. Similarly for “I know Sam”—in the sense of “I am acquainted with Sam”—in order for me to know Sam in this sort of way, Sam must exist.

Putnam claims that in “traditional philosophical psychology” it has been assumed that the vast bulk of mental states are narrow, and hence presuppose the existence of nothing beyond the particular subject to which they belong

(1975, p. 220). This view of the mental is certainly implicit in presentations of Cartesian skepticism that assumes that one's mind could be just as it is even in the absence of an external world. Putnam is careful to make it clear that in saying that it is possible for subjects to be in the same psychological state but mean different things by (say) "water," he is referring only to psychological states in the narrow sense. His goal, in effect, is to get us to appreciate that *understanding the meaning of a word* should itself be viewed as being a broad sort of psychological state, not a narrow one—this is where previous theories of meaning have gone wrong, by supposing that understanding is wholly a matter of what goes on in one's head (or mind).

However, it was not long before some philosophers—for example, McGinn (1977)—were arguing that the Twin-Earth considerations had wider and more profound implications for the nature of the mental than Putnam himself realized—and Putnam himself later acknowledged as much. It is not difficult to see why. A substantial part of a person's mind is made up of their propositional attitudes—their beliefs, hopes, fears, etc.—and in specifying what a person's propositional attitudes are we use "that clauses," such as "Sam believes that water is wet," or "Mary thinks that water is horrid," and so on. Let us suppose Putnam is correct, and that "water" on Earth does not mean the same as "water" on Twin Earth. Does it not quickly follow that all the many "water"-related beliefs (and other attitudes) that you and your doppelgänger possess are going to be significantly different in character? You believe that water is wet, and so does your doppelgänger on Twin Earth. But these beliefs are different: the belief that water is wet is distinct from the belief that the Twin-Earth water is wet. Why? Because the beliefs have different *contents*: one is about H₂O, the other is about XYZ. This point generalizes: think of all the people you know, and all the people your Twin-Earth doppelgänger knows. These are all different people, living as they do on different planets. So your belief that "Bismarck was a politician of god-like powers" is not the same belief as your doppelgänger's belief that "Bismarck was a politician of god-like powers," and for analogous reasons it is clear that a great number of your propositional attitudes are going to be distinct from those of your doppelgänger. If this is indeed the case, the implications are obvious, and radical: to a very considerable degree our minds are *not* independent of the surrounding world. People who are physically indistinguishable so far as their brains are concerned—so far as their internal states are concerned—can be mentally very different. In which case, the traditional view of the mind as something self-contained, contained within the body (or soul) is very much misguided.

This general line of argument has since become known as *externalism*. Not all externalists are equally radical. Those who are more radical hold that just about all forms of mental content are broad—in Putnam's sense—and that this holds for both propositional attitude states and content-bearing

experiences (such as perceptual states). Some contemporary externalists—such as McDowell (1986)—defend a variant of Russell’s view that some mental states have the objects they are about as constituent parts. The opposing doctrine of *internalism* also comes in different strengths, with some accepting that there is both broad and narrow content, while others hold that the only sort of mental content is content of the narrow variety; there are also different conceptions of how narrow content should be conceived. The debates between and within these competing camps are one of the primary points of intersection between the philosophy of mind and the philosophy of language in recent years, and are likely to remain so for the foreseeable future.⁹

15 Analytic Philosophy of Mind

Barry Dainton and Howard Robinson

Philosophy of mind is one of the dominant fields in analytical philosophy nowadays, but this was not always or evenly so. Most of what we have described in our history of the tradition relates closely, in one way or another, to the philosophy of language, formally or informally conceived. Some strains in the philosophy of mind relate to this emphasis and others do not.

The central concern in most philosophy of mind has been “the mind-body problem,” and the emphasis has been to see how far one can go in providing a *naturalistic*—that is, materialist and natural-science oriented—account of the mind, which mainly pertains to consciousness and thought. It would probably be fair to say that this issue came to be focused on as a result of the apparent success of mechanistic science in the latter half of the nineteenth century. The question arose as to what, if the physical world—including the human body—is a deterministic machine, are we to do with consciousness: how can we find a place for it in the physical universe? The answer made popular by T. H. Huxley and others was that consciousness is an epiphenomenon—a byproduct of the physical machinery, which does not influence it in return. In this way one preserves the integrity of the physical machine, while acknowledging the existence of consciousness as an extra phenomenon. But epiphenomenalism is a desperate theory. The idea that the pain that I feel when hit or the visual experience I have of seeing the lion bounding toward me, have nothing to do with the way I react, can only be held as a last resort in order to save a theory. If one wishes to avoid such a bizarre position, then one seems to face the choice between a return to dualist interactionism, thus apparently abandoning a belief in the integrity of physical science, or finding some way of building the mind into the physical system. The main efforts of analytical philosophers went in the latter direction.

The philosophical urge to find a naturalistic solution to the mind-body problem coincided with a move in psychology to make the subject more rigorous and “scientific.” The latter meant moving from relying on the introspective

reports of subjects (as in, for example, the psychology of William James) to something more “measurable.” This supposedly measurable quantity was *behavior*, and, to make it simpler, the behavior, not mainly of humans, but, in the first instance at least, of pigeons. This, in the hands of Watson, gave rise to behaviorist psychology.

The core or generic idea behind the behavioristic approach to the nature of mind is that minds and their contents consist entirely in the contribution they make, directly or indirectly, to outward and in principle observable behavior. What is traditionally supposed to be the “inner life” of the mind is declared to have no more to it than the difference it potentially makes to outward behavior.

In its basic form, behavioristic psychology consists in trying to establish laws linking external stimuli to overt responses. Except for very crude cases, such as pain reactions or the simplest learning by animals, this can cover very little of psychology. But there developed a whole theory of learning, involving conditioning and reinforcement by external stimuli, which, it was hoped, would enable behaviorism to cope with more complex phenomena. B. F. Skinner is the best-known psychologist in this tradition and Quine probably his best-known philosophical follower. But the first major uptake of behaviorism by philosophy was not in this exclusively external stimulus, external response form. The logical positivists, believing as they did in the reduction of everything to a “unified science,” wanted to reduce psychology to physics—as we saw in Chapter 8—which means to any physical quantity that could be measured. Behavior is included among such physical quantities, but so are other physical events, such as ones occurring in the brain (see Hempel 1949). Sometimes inner physical events, such as neural events or changes in blood-pressure, are characterized as “covert behavior.” Of course, in the normal sense they are not *behavior* at all—they are not *actions* that we *do*. But they are observable physical events and so can be included in a stretched sense of the term. This use of the term points to an important ambiguity in “behavior” in this context. The physicalist’s main ambition is to show that one need not postulate an inner, private, immaterial realm of conscious experience. Characterizing mental states in behavioral terms achieves this. A second ambition, required by the positivist program of a unified science, is to give an account of psychological predicates in nonpsychological terms.

One could serve the first purpose by explaining pain, for example, as the disposition to *pain*-behavior, where the identification of the behavior is dependent on the experiential notion, *pain*. Similarly, one might talk of belief behavior, or verbal behavior, leaving these dependent on the notions of belief and language. The second ambition, on the other hand, would require that the behavior in question be characterized in simple physical terms—as, for example, movements or sounds of certain physical types, the descriptions of which

required no psychological predicates. It was soon realized that this latter ambition is completely impossible. Even Skinner's behaviorism involved attributing to the pigeons that peck in order to receive food, the *desire for* or *purpose of achieving* food. In Ryle's language, mental dispositions are *multi-track*: that is, a particular mental state could be realized by many forms of physical behavior that are unified as a kind only because they subtend the mental purpose in question and hence can only be characterized in mental terms. Even a state with relatively straightforward relations to behavior, like pain, has many physically varied manifestations—saying “ouch,” screaming, rubbing the damaged spot, going to the medicine cupboard, etc. Without the concept of pain there could be no reason for classifying these behaviors together.

How far does it matter to the physicalist if psychological states cannot be type-characterized in purely nonpsychological terms? This is closely connected to the issue of whether a physicalist must be a *reductionist* or whether there is such a thing as *nonreductive physicalism*. As we shall see, this issue itself is very murky.

For most of the period since about 1970, physicalists have wanted to claim that they are not reductionist, but the issue has been rendered unclear by lack of a consistent use of the term. Philosophers with a background in the philosophy of science tend to take the term in the sense given it by Nagel in his *The Structure of Science* (1961). There, reductionism involves type identity for properties and natural laws. So water can be reduced to H_2O because all and only water is H_2O and all and only the laws that apply to water apply to H_2O . By this standard, functionalists (we will be looking at their position shortly) and most forms of behaviorist are not reductionists, because functional states and behavioral dispositions are *multiply realizable*, which means that, like a computer program they can, in principle, be built into different kinds of hardware. All water is H_2O but not every accelerator or armchair is made of any given material: this is, in general, true for all functional and teleological concepts.

On the other hand, for many philosophers, functionalism and behaviorism are paradigmatically reductionist theories, for they claim that mentality is “nothing but” something that contributes to behavior, and not the special, subjective, private, quasi-Cartesian phenomenon that it seems to be. There is here, a difference between British and Australian philosophers, on the one hand, and American ones on the other. The behavioristically inclined among the former generally took their anti-dualistic approach to mind from Ryle and Wittgenstein and not, like their American colleagues, from positivist philosophers of science who had escaped from Central Europe. Consequently the British and Australians were never greatly influenced by rigid scientific reductionism. The idea that psychological vocabulary might be autonomous, serving a different purpose from the vocabulary of the physical sciences, did not worry them, so long as a broadly sociobehavioral gloss could be put on how

such concepts worked. Resistance to the behavioristic approach did not come from within the philosophy of science, but from the intuitive feeling that any purely behavioral approach left something out, namely the raw data of bare experience. This was what led U. T. Place, J. J. C. Smart, and D. M. Armstrong to augment their behavioral approach with central state materialism, identifying mental states not with the rather abstract entity, a *disposition* to behave, but with the solid inner machinery that caused the behavior. This causal theory of the mind tied up naturally with the functionalism that will emerge below. For those British and Australian philosophers who did not like a materialist approach to the mind, Ryle, Wittgenstein, Place, Smart, and Armstrong would all count as reductionists, for they thought that mind was *nothing but* what is manifested in behavior, even though there was no scientific account of reduction involved.

Because of these different approaches to what being a reductionist meant, when Donald Davidson proclaimed himself to be a nonreductive physicalist, in his famous paper "Mental Events" (1970), his meaning in so describing himself was diversely interpreted. His proclaimed ground for claiming to be nonreductionist was his denial that there are any psychophysical laws, but his theory was taken up as if it were something distinctively less reductive than functionalism, even though they also denied the existence of such laws (see Robinson 2002).

It is not possible to go further here into Davidson's theory, but it is important that he imported the term "supervenience" into the philosophy of mind. One state or kind of state *supervenes* on another state or kind of state if the former cannot vary independently of the latter. So if the mind supervenes on the brain, then one cannot have a change of mental state without a change of physical state. Supervenience is meant to be less ambitious than reduction, as defined by Nagel, because the dependence is one way only; there could be a change in brain state that did not involve a change in mental state (e.g. if the change was only a small one). In the eyes of its critics, the problem with the deployment of the concept of supervenience in the philosophy of mind is that it is the name for a putative phenomenon masquerading as an explanation of it. If we were to agree that the mind cannot change without the brain changing, we will still want to know why this is so—why can the mind not vary independently? Traditional identity theory and strong reductionism provide answers. It is, roughly, because they are the same thing. But simply to stipulate that they cannot has no explanatory force. One might say that mind cannot vary without brain variation because mind is causally dependent on the brain, but such mere causal dependence is consistent with property and maybe even substance dualism. So the idea becomes that supervenience is a more-than-causal necessary dependence, but less than identity or strict reduction. Simply calling such a supposed dependence "supervenience," or a *sui generis* brand

of “metaphysical necessity,” does nothing to make the idea plausible or even contentful.

The strict stimulus-response behaviorism of a Skinnerian kind came under severe attack in the late 1950s. The brunt of the attack concerned its emphasis on the role of external stimuli in controlling our behavior and the denial that internal processes did more than transmit the influence of the stimulus to the response. Chomsky’s criticism of Skinner on the topic of language learning (1959) and the rise of cognitive science, both of which stress the importance of inner processes and innate capacities, forced stimulus-centered behaviorism into retreat and saw the rise of *functionalist* theories of mind, which emphasize the distinctive role of the “computational program” in the head. Putnam’s “machine state” functionalism (1960, 1967) was an early example. According to Putnam, the mind of any minded entity—humans included—can be regarded as being akin to a computer (or Turing machine), and the program (or “machine table”) for this computer gives a complete account of its mental states and mental structure. The states characterized by such a program differ from the simpler dispositions posited by the behaviorists. They are not characterized solely in terms of their relationship to perceptual stimuli and behavioral responses; instead their characterization also makes essential reference to the specific ways they can influence—perhaps only probabilistically—other states within the same system, and be influenced by them. This more sophisticated model of the mind—as a system of interdependent, causally interacting states—quickly found favor in many quarters. Moreover, the functionalist approach was significantly boosted when David Lewis (1972), drawing on earlier work of Ramsey’s, proposed a way of formulating functionalist theories that was free from any hint of circularity—Lewis’ method makes it possible to exclude, in rigorous fashion, any reference to the mental from functionalist-style definitions of mental states.

Nevertheless, despite its sophistication, in the eyes of its critics functionalism is a behavioristic philosophy of mind, because the mental or psychological significance of the inner workings consists entirely in the contribution they are primed to make ultimately to behavior. In David Armstrong’s words, “the mind is not behaviour, it is the cause of behaviour.” Philosophical behaviorists such as Ryle identified mental states with *dispositions* to behave, but whether a disposition is actualized at a given time does not depend only on how one is stimulated but on what other dispositions—including other mental states—one has. Whether one responds in a given way to a certain stimulus—a pain, or a question or a problem posed—will depend on a massively complicated mental background, that is, on all the other relevant dispositions one possesses. This is so whether one thinks of this background as all derived from earlier stimuli, whose influence has been stored over the

previous years of one's life, or from innate cognitive capacities. The mental models that the functionalist builds of the inner structure of our minds can be thought of as a way of bringing some organization into our understanding of this mass of acquired or innate dispositions. From a philosophical point of view, functionalism can be regarded as behaviorism with a complex model of how one gets from stimulus to response, given all the other stored dispositions one possesses.

The essential similarity of behaviorism and functionalism from a philosophical point of view is shown by the fact that the major intuitive objection to both is the same, namely that they cannot cope with the subjectivity, the "what it's like" or "raw feel" of experience. This is because they both try to export the inner, subjective life into the realm of the publicly observable. This line of attack was first brought into focus by Thomas Nagel (1974)—no relative of Ernest Nagel, the definer of reduction—given greater currency by Jackson (1982) and further elaborated by Chalmers (1996). Nagel claimed that, however much we came to know about the echo-locating mechanism of bats, we, as humans lacking such a sense, could never know *what it was like* to be a bat using this faculty. Jackson imagined a scientist, Mary, who knew everything there was to know about the physics and neurology of normal vision, but who herself was carefully protected from experiencing any chromatic colors. If she were finally allowed to see colored objects, she would discover something new, namely, *what it is like* to see chromatic color. These arguments are meant to show that knowing all the physical facts still leaves something unknown, namely the subjective, *what it is like* or qualitative nature (qualia) of experience. Hence the physical facts are not all the facts and physicalism is false. This is known as "the knowledge argument," and it has been responsible for a very considerable amount of discussion in recent years.

15.1 More Recent Work

Throughout the 1970s and 1980s there was considerable enthusiasm for a variety of new (and sometimes not so new) topics, approaches, and theories, many inspired by scientific developments. We will confine ourselves to mentioning just some of the most important—or intriguing—examples.

The work of Dretske (1981) and Evans (1982) triggered a still-vigorous debate on the existence of "nonconceptual content." Those who follow Dretske and Evans in defending nonconceptual content—see Peacocke 1986, 1989, 1992, Cussins 1990, Crane 1992, Tye 1995, 2000—hold that mental states can represent in nonconceptual ways. If so, then the content of some perceptual states, for example, are far less like the contents of thoughts or beliefs than would otherwise be the case. Critics, such as McDowell (1994) argue that

perceptual experience would not be able to rationally justify beliefs unless they both possess conceptual concepts.

As we have just seen, thanks in part to the influence of Davidson, in the 1980s there was much debate relating to the viability of “nonreductive” (or supervenience-based) forms of physicalism. Whether or not this way of conceiving of mental states was compatible with the latter retaining causal powers was (and is) a particularly contentious issue (for more on this see Beebe, this volume, section 5.2). Functionalism remained an influential general account of the nature of the mind, and much work was done on refining and elaborating on the basic thesis.¹

Although functionalism was itself—in part—inspired by the development of computers, the latter also impacted on cognitive science, and developments there—for example, new ways of thinking about the concepts, computational models of visual processing, particularly Marr’s (1983) contribution—quickly made an impact on philosophical debates too. According to Fodor’s influential “representational theory of mind” (RTM), mental processes are computational processes, involving syntactically structured representations; a human mind is thus a “syntax-driven machine” in precisely the same manner as a digital computer—see Fodor 1975, 1987.² If the mind is furnished with representations, finding a viable naturalistic account of mental representation was (and remains) a pressing issue for physicalist philosophers (Fodor 1981, Block 1986). In his much-discussed “Chinese Room” argument, Searle (1980) argues that syntactic manipulations alone can never give rise to genuine understanding.

The rise of the connectionist movement in the late 1980s challenged the then-dominant assumption in cognitive science circles that human cognition is closely akin to classical symbolic computation at the heart of RTM: connectionist computers—also known as artificial “neural nets,” or “parallel distributed processing” devices—are nothing like digital computers, and do not engage in symbol manipulation. There ensued a lively—and still continuing—debate about the contrasting merits of these different modes of computation, and their relevance to human cognition.

The 1980s also saw the rise of “eliminative materialism.” On this view—see, for example, Churchland 1981, Dennett 1988—our ordinary ways of thinking about our minds are highly likely to be supplanted by future developments in psychology and neuroscience. We no longer think that the Earth is at the center of the universe, or that falling objects are attempting to return to their natural place: these primitive and erroneous doctrines have been eliminated by advances in natural science. In a precisely similar way, as and when science discovers more about the brain, our primitive ways of thinking about the mind will almost certainly be shown to be drastically mistaken too. The notion that there are such things as *beliefs*, *hopes*, *desires*, and *fears* will strike our descendants as faintly ludicrous.

The philosophy of mind may have been flourishing as never before, but by the 1990s a small but growing number of philosophers gave voice to the suspicion that the then dominant naturalistically oriented accounts of mentality all failed to do justice to one aspect of the mind: consciousness. Nagel had argued as much in his "What Is it Like to Be a Bat?" (1974), and a few years later Levine (1983) argued that there was an "explanatory gap" separating the physical and the experiential. Despite their other differences, the failure of extant versions of physicalism was common ground for McGinn in *The Problem of Consciousness* (1991), Searle in his *The Rediscovery of the Mind* (1992), and Galen Strawson's *Mental Reality* (1994).³ Of these McGinn was by some measure the most pessimistic: he argued that not only was reductive materialism doomed to failure, the problem of understanding how the brain produced experience was very likely too difficult for us ever to solve. Searle was notably less pessimistic, arguing that subjectivity is a natural, biological phenomenon, albeit an emergent one. Strawson defends a "naturalized Cartesianism," arguing that if materialism is true, then we are radically ignorant of the real nature of the physical.⁴

In 1996 David Chalmers' *The Conscious Mind* was published. The book soon had a significant impact and was widely discussed beyond the confines of philosophy. It is also noteworthy for the way in which it uses the technical resources distinctive to analytic philosophy—in this case the "two-dimensional" approach to modality—as part of a defense of a position that hitherto had largely been ridiculed in analytical circles: a dualistic account of the mind. Chalmers starts by distinguishing between the "easy" problems in the philosophy of mind, those that look as though they can be solved by the standard methods of cognitive science, from the "hard" problem of *experience*: how and why does the physical activity in our brains give rise to consciousness? Chalmers goes on to argue that since none of the standard reductive accounts succeed, the hard problem remains unsolved, and so we need to consider more radical alternatives.

An important component of Chalmers' argument against reductive physicalism is the "zombie" argument. We can, he argues, conceive of a being who is exactly similar to a normal human in all physical respects, but who experiences nothing at all—a being who is a zombie (in the philosophical, rather than Hollywood, sense). If such a being is possible, experience cannot logically depend (or supervene) on the physical, in the way that physicalists hold. But can we conclude anything about real (metaphysical) possibility from what we are able to imagine or conceive? Many are skeptical of this. Indeed, one of the most significant developments in the 1970s was Kripke's demonstration that there can be necessary truths that are a posteriori. If "water is H₂O" is necessary but also a posteriori as Kripke argues, then it is conceivable that water is *not* in fact H₂O, even though this is not in fact possible. This widely accepted

Kripkean line of argument breaks the connection between what is conceivable and what is genuinely possible. In response, Chalmers opts for a particular interpretation of two-dimensional modal semantics, a way of thinking about necessity and possibility that retains a connection between conceivability and possibility. Within this framework, or so Chalmers goes on to argue, there is a sense in which what is conceivable *is* logically possible.⁵ Furthermore, since zombies fall into this category, we should conclude that physicalism, at least of the reductive variety, is false.

If physicalism is false, what should we conclude about the place of consciousness in the world? There are many options, but Chalmers suggests that the most promising is a “naturalistic dualism.” We should take experience to be a *fundamental* ingredient of reality, just as fundamental as mass or charge; but whereas the latter properties are physical, experiential properties are non-physical. Two factors render Chalmers’ version of property dualism naturalistic. First, he holds that experiences will be bound to the physical realm by a web of natural psychophysical laws linking the distribution of experience to patterns of computational activity. Second, these psychophysical laws will not in any way interfere with ordinary physical laws—Chalmers is fully prepared to accept that the physical is causally closed.

This naturalistic dualism has its advantages, but it also comes at a significant cost: since physical events only have other physical events as causes, it renders the experiential epiphenomenal. But there is, as Chalmers acknowledges, an alternative. Experience has no place in the physical world if the properties of the latter are confined to the sort found in fundamental physics, and the various higher-level properties—such as liquidity or solidity to which these can give rise. But perhaps physics is not the whole story when it comes to the physical. Over and above the causal and structural properties of physical things that feature in our scientific theories, perhaps there are intrinsic properties of a sort that are not recognized by these theories. Our current best theories relating to the elementary particles, for example, tell us a great deal about how the various types of particles interact with each other, but they tell us nothing at all about the *kind of stuff* that possesses these causal properties. If we suppose that the basic forms of matter have properties over and above their causal and structural features, then if some of these additional properties are experiential in nature, we have found a way in which experience can be a part of the physical world.

This form of materialism is the option favored by Lockwood (1989), Strawson (1994), and Stoljar (2001)—also see Stoljar’s contribution to this volume. Something very much like this view had long been held by Russell: “The gulf between percepts and physics is not a gulf as regards intrinsic quality, for we know nothing of the intrinsic quality of the physical world, and therefore

do not know whether it is, or is not, very different from that of percepts" (1927/1954, p. 264). And as a consequence in the contemporary literature it is often referred to as "Russellian Monism."⁶ What is less often remarked upon is that by incorporating experience into the most elementary of material things, this position is not so very distant from the idealism that Russell had rebelled against only a couple of decades earlier.

Appendix

A Simple Introduction to Tarski's Theory of Truth

Barry Dainton

1 Some Basic Ideas

The starting point is the *Equivalence Thesis*, according to which for any sentence such as "It is true that zebras are striped," the "It is true that" clause can be removed without a significant loss or change of meaning, so that our original sentence is equivalent to simply "Zebras are striped." More generally, to say that a sentence "p" is true is to express no more than p all by itself. Abbreviating "if and only if" to "iff," we can express this equivalence by saying that:

"p" is true iff p

Here the sentence p on the right-hand side tells us the necessary and sufficient conditions for the truth of the left-hand side sentence: snow's being white is necessary and sufficient for the *truth* of "Snow is white." One of Tarski's key ideas is that we can use this equivalence, relying upon the "transparency"—in effect, the near-redundancy—of the truth-predicate to generate a "materially adequate" definition of the term "true sentence" for a specific language, that is, a definition that will accurately specify the conditions in which all the relevant sentences are in fact true.

The first step was to distinguish the "object language," the language we are defining truth for, and the "metalanguage," the language we are formulating the definition in. Given this distinction we get so-called *T-sentences* such as:

(T) "s" is true-in-L iff p

Here, p stands for a sentence in the metalanguage that has the same meaning as "s." The sentence "s" is not *used*, but merely mentioned or referred to,

whereas the sentence *p* is used—it has its ordinary meaning. Consequently, it follows from the equivalence thesis that the T-sentence gives us the necessary and sufficient conditions for the truth of the sentence “*s*.”

Tarski argues that each T-sentence for a language *L* can be regarded as supplying a *partial definition* of the term “true sentence” for language *L*. For each T-sentence tells us the necessary and sufficient conditions for the truth of a particular sentence of *L*. Consequently, if we could specify *all* the T-sentences for *L*, we would (says Tarski) have provided a complete definition of “true sentence” for *L*. Or at least, we would have a materially adequate definition, one that tells us, for any sentence in *L*, the conditions under which that sentence is true.

With these fundamentals out of the way, we can take a more detailed look at his theory.¹

2 A Truth-definition for a Basic (Non-quantified) Language

Our simple artificial language contains only a limited number of names and a limited number of predicates (there are no logical connectives any sort), which together allow it to express a limited number of sentences. The syntax of our simple language is very simple indeed: if “*P*” is a predicate and “*a*” a name, then putting them together thus: “*Pa*” gives a well-formed sentence. The semantics is straightforward too: names refer to objects, predicates stand for properties, and any sentence says of the thing named that it possesses the property referred to by the predicate (or “satisfies” the principle for grouping objects introduced by the predicate). This can be expressed by a *compositional axiom* of this form:

- (D) Any sentence *Pa* in any language of this sort will be true iff the predicate applies to, or is satisfied by, whatever it is that the name refers to.

(D) tells us that the truth of a sentence in our simple language is a matter of two factors: one concerning the reference of a name, the other the satisfaction of a predicate. It is worth noting that (D) does not tell us when a sign is being used *as* a name or predicate, how objects are named, or what determines whether an object named satisfies a predicate. We are simply taking it for granted that some names *do* refer, and some predicates are satisfied by objects.

Thus far we only have a very general framework for the semantic description of our simple language, but nothing more. To get a description of a *specific* artificial language we need to specify what names and predicates it contains,

and what they stand for. Let us consider two artificial languages, L_1 and L_2 , each containing two names (a , b) and two predicates (P_1 , P_2):

a in L_1 refers to Lenin
 b in L_1 refers to Marx
 P_1 in L_1 applies to bald things
 P_2 in L_1 applies to pink things

 a in L_2 refers to Paris
 b in L_2 refers to Rome
 P_1 in L_2 applies to French things
 P_2 in L_2 applies to warm things

Each language can express four sentences. The framework (D) and these interpretations can be called the *applied framework*, and using the latter we can work out the truth-conditions of each of these sentences in each language: for example, that " P_1a " is true in L_1 iff Lenin is bald, and in L_2 iff Paris is French.

Note: these languages are wholly artificial, in that we have simply *stipulated by fiat* the meanings of their various terms. Although they could be spoken by some actual people, in specifying their semantics in this way we have not said anything about what would have to be true of such speakers in order for this to be the case (e.g. what would have to be true of their psychology, conventions, behavior), or how we would empirically establish that this was the case.

3 Truth-in-L and Convention-T

We could view the applied-framework for L_1 as a *manual* allowing us to compute what has to obtain for any sentence in L_1 to be true. Since the manual is complete, it could be said to "characterize" this *language*. However, Tarski preferred to say instead that the manual characterized "definitions" for truth, or reference, or satisfaction, or the language. In so doing, he took himself to be establishing a scientific semantics, by showing how to replace semantic terms altogether for suitably simple languages.

Tarski takes the clauses in the applied framework (e.g. that " a " in L_1 refers to Lenin, etc.) to be *partial definitions* of the associated semantic concepts, the *full* definitions being given by the total set of these partial definitions and their consequences. For L_1 (and likewise for L_2), we can view these clauses as partial definitions of these hyphenated concepts: reference-in- L_1 , satisfaction-in- L_1 , truth-in- L_1 , where:

X refers-in- L_1 to Y iff: X is a and Y is Lenin, or X is b , and Y is Marx

Y satisfies-in- L_1 X iff: X is P_1 and Y is bald, or X is P_2 and Y is pink

S is **true-in- L_1** iff: S is P_1a and Lenin is bald; S is P_2a and Lenin is pink; S is P_1b and Marx is bald; S is P_2b and Marx is pink.

We are meant to think of the right-hand sides of the definitions as full and complete definitions of the hyphenated terms, with each item in the list being *part* of the definition.

Moreover, note that we have here a definition of truth-in- L_1 that does not employ *any semantic terms at all*. Using the list we can tell whether any sentence in L_1 is true without using any semantic judgments: all you need to be able to do is tell which sentence you are dealing with (which you can do without knowing its *meaning*; its spelling will suffice) and tell (for example) whether Lenin is pink.

Our simple languages enable us to give “list-accounts” of true-in. . . refers-in. . . satisfies-in. . . since there are only a finite number of sentences formulable in either L_1 or L_2 . No list-type account of this sort would be possible if we could construct an infinite number of sentences; if this were possible we would need a *recursive clause* to specify “true-in. . .” that is, a clause that would tell us how to add an appropriate sentence to the list for “true-in. . .” (and the rest) for any sentence in L we were given, no matter how complex.

This is not difficult to do. Suppose we enrich L_1 by adding the connective “&” (meaning conjunction). L_1 can now express an infinite number of sentences: for we can conjoin any two to form a new sentence, and then conjoin this new sentence with one of the originals to form yet another sentence, and so on indefinitely. But the definition of true-in- L_1 is easily amended to take this into account:

“P and Q” are true-in- L_1 iff P is true-in- L_1 and Q is true-in- L_1

By applying this rule to any conjunction in L_1 , we will eventually reduce the question of its truth-in- L_1 down to a series of questions about the truth-in- L_1 of the basic four sentences (as exemplified for the simplest case in the statement of the new rule itself).

Tarski’s impressive technical achievement was to show how definitions of just this sort could be given for more complex languages containing quantification.

4 Truth-in-a-quantified Language

For the more complex language we will be considering next, we need the basic apparatus of the predicate calculus.

Accordingly, we will let capitals such as F, G, H, stand for predicates; lower-case letters at the start of the alphabet such as a, b, c . . . stand for objects; lower-case letters toward the end of the alphabet, such as x, y, z, are *variables*, which when conjoined with a predicate, as in “Fx” form “open sentences,” that is, “sentential functions” that produce sentences when a name is substituted for their variables. Open sentences are not proper sentences in their own right; however they become so when preceded by a quantifier: $(\forall x)Fx$ means “everything is F” $(\exists x)Fx$ means “something is F.” Variables in open sentences are “unbound,” variables in quantified sentences are “bound.”

In this language there are two sorts of proper (or “closed”) sentences: simple atomic sentences like “Fa,” meaning the object referred to by “a” has the property “F-ness,” and quantified sentences, such as $(\forall x)(Fx \rightarrow Gx)$, meaning “for any object x, if x is F then x is G,” or in other words, “all F’s are G’s”). Or again, $(\exists x)(Fx \& Gx)$, meaning “there is some x, which is F and which is G,” or in other words, “something is both F and G.”

5 A Problem, and a Strategy for Solving it

In the simple quantifier-free languages (such as L_1 and L_2), we could define the truth of compound sentences ($X \& Y$) in terms of the truth of (satisfaction of predicates by objects) their parts, for these parts are themselves *sentences* in their own right. We cannot do this in a quantified language, because the component parts of typical quantified expressions, such as $(\forall x)(Fx \rightarrow Gx)$, are not (closed) sentences, but open ones.

Tarski’s solution to this problem was to axiomatically (recursively) define a more general notion than truth, *satisfaction*, which is applicable to open sentences and predicates. He would then go on to define truth in terms of satisfaction.

6 Defining Satisfaction

The basic relationship is between *sequences* of objects on the one hand, and predicates and variables on the other. The things that do the “satisfying,” the “satisfiers,” are sequences of objects; the things that are satisfied are predicates with one or more variables. A “sequence” of objects is a collection of objects in a particular *order*: the same objects ordered in different ways make different sequences.

The basic idea can be illustrated thus. An open sentence such as Fx is satisfied by a particular sequence of objects if the *first* member of the sequence is F (e.g. “is funny”). A two-place open sentence such as Hxy is satisfied by a particular sequence if the first member is H-related to the second (e.g. “. . .

is happier than . . ."). And so on. All the other remaining members of the sequence are irrelevant. Since we assume that there is no limit to the number of variables that a sentence can possess (we can make sentences as complex as we like), the basic account of satisfaction deals with sequences with an infinite number of objects (the same object can occur in the same sequence as many times as we like).

To ensure that we have an infinite supply of different variables, we "index" them from one to a potential infinity ($x, x'', x''' \dots$), and assign the first object of the sequence to the variable possessing the first index, the second object in the sequence to the variables possessing the second index, and so on. There is thus a one-one correspondence between each variable occurring in a sentence and objects occupying distinct places in the sequence we are concerned with. The existence-conditions of sequences are specified thus:

- (a) There is at least one denumerable (countable) sequence.
- (b) For every sequence S , for every natural number i , and for every individual y , there is a sequence S' which differs from S in at most the i -th place, and whose i -th member is y .

This second condition may seem odd, to put it mildly. But it boils down to this: for any given sequence S , such as $\langle \text{Tarski, Quine, Russell, . . .} \rangle$ there are other sequences that differ from S in just one respect, that is, for each position in S , instead of the object that is actually there, *any* other (different) object could be there. Focusing just on the second position in S ($i=2$), some possible sequences are $\langle \text{Tarski, Spinoza, Russell, . . .} \rangle$, $\langle \text{Tarski, the Eiffel Tower, Russell, . . .} \rangle$, $\langle \text{Tarski, Russell, Russell, . . .} \rangle$, and so on (as you see, the same object can occur more than once). Any object other than Quine could go in the second position. And the same goes for each other location in the series.

Obviously, given one sequence S , the second condition guarantees the existence of many others, in the form of every possible single-change permutation of the original S .

From now on in it is plain sailing.

7 A Truth-definition for QL

We start, as previously, with names and predicates.

- (a) Names
 - (i) "a" refers to a
 - (ii) "b" refers to b
 - (iii) . . .

(b) Predicates

- (i) An object a satisfies "F" iff a is F
- (ii) An object a satisfies "G" iff a is G
- (iii) . . .

We next specify the conditions under which the atomic formulae of the language are satisfied:

- (c) A sequence S satisfies a predicate F concatenated with a name n iff the object to which the name refers satisfies F .
- (d) A sequence S satisfies a predicate F concatenated with the k -th variable (v_k) iff the k -th member of the sequence (s_k) satisfies F .
- (e) A sequence S satisfies the formula " $(F \ \& \ G)$ " iff it satisfies F and it satisfies G .
- (f) A sequence S satisfies the existential quantification of a formula A with respect to v_k iff A is satisfied by *some* sequence S' that is like S except perhaps in the k -th term.
- (g) A sequence S satisfies the universal quantification of a formula A with respect to v_k iff A is satisfied by *all* sequences S' that are like S except perhaps in their k -th term.
- (h) A sentence is true iff it is satisfied by all sequences.

8 How Do the Definitions Work?

At first view they can look utterly baffling. In fact, they are not so difficult, but there is a fair amount of formal maneuvering involved.

Consider (c): we are concerned here with closed sentences, such as "Fa" ("Alan is funny," say). Since satisfaction is concerned only with open sentences and variables, we can say what we like here: we can say a given true atomic sentence is satisfied by all sequences or by none. It is convenient to say that all true atomic name-predicate sentences satisfy ALL sequences. So this is what we say.

Now consider (d): we are concerned here with formulas containing unbound variables. We can focus on " Fx " the index indicating that this is the second variable in our list. Suppose " F " is the predicate "is a philosopher," then the formula is satisfied by the sequence $\langle \text{Hitler, Quine, the moon, } \dots \rangle$ but not by $\langle \text{Hitler, the moon, Quine, } \dots \rangle$ since the object in the second position is not a philosopher. Note: that the satisfaction conditions for open sentences like this are not "all-or-nothing" as they are for closed sentences, that is, it is not the case that a formula is satisfied either by all sequences or none. But since we are not dealing with *sentences* here, this does not matter.

The satisfaction condition for the conjunction of such formulas, (e), as well as the other sentence connectives is trivial. So we pass onto (f) and (g), the quantified sentence conditions, and see why they yield all-or-nothing satisfaction conditions.

In the *existential case*, (f), satisfaction conditions are defined thus:

- (f) A sequence S satisfies the existential quantification of a formula A with respect to v_k iff A is satisfied by *some* sequence S' that is like S except perhaps in the k -th term.

Suppose we are dealing with the sentence "Someone cuts my lawn and prunes my roses," which is symbolized thus:

$$(\forall x'')(Fx'' \& Gx'')$$

We have chosen (arbitrarily) the second indexed variable here (there are others, any of which would have done, such as x , y , y'' , z , etc.). Suppose the sentence is true, because someone, Tom Brown, did cut by lawn and prune my roses. Now consider any sequence such as:

$S \langle \text{Cicero, the moon, Quine, Russell, } \dots \rangle$

This does not satisfy the sentence, since the second-place object did not cut my lawn. On the other hand, this sequence does satisfy:

$S' \langle \text{Cicero, Tom Brown, Quine, Russell, } \dots \rangle$

Assuming this sequence continues just like S for the remainder of its extent, then S' differs from S in just the second position, where lies Tom Brown in place of the moon. So S' , unlike S , satisfies the open sentence " $Fx'' \& Gx''$," which is what our existential quantifier quantifies. Recall now what the definition (f) tells us: if an open sentence (A) is satisfied by a sequence S' differing from S in at most one position, then S also counts as satisfying the existential quantification of that formula. (Note that the satisfying sequence S' must differ from S in *at most* one place; it need not differ at all, in which case $S = S'$). It immediately follows that ALL sequences will satisfy the sentence, since any that do not actually satisfy it can be turned into ones that do by changing only one of its members (substituting Tom Brown for whatever is at the second position).

But what if the sentence is false because no one cut my lawn and pruned my roses? Clearly, in this case no sequence will satisfy the relevant open sentence. How could it if the required object does not exist (and so is found in no sequence)?

So, the definition guarantees all-or-nothing satisfaction conditions for existentially quantified sentences: if just one sequence satisfies the relevant open sentence, all will. All this may seem to have been accomplished in a trivial manner, but it does the job.

In the *universal case* the relevant definition is:

- (g) A sequence S satisfies the universal quantification of a formula A with respect to v_k iff A is satisfied by *all* sequences S' that are like S except perhaps in their k -th term.

On the face of it this might seem more problematic, since A has to be satisfied by ALL sequences S' rather than just SOME as in the existential case. But this matters not. Consider the sentence:

All aardvarks are flea-ridden.

Symbolized as:

$$(\forall x'')(Fx'' \rightarrow Gx'')$$

From elementary propositional logic, this is equivalent to:

$$(\forall x'')(\sim Fx'' \vee Gx'')$$

Now consider the sequence:

$S \langle \text{Cicero, Quine, the moon, } \dots \rangle$

For satisfaction, what is required is both that the second member of this sequence satisfies the open sentence $\sim Fx'' \vee Gx''$, and also that the second member of all sequences S' differing from S in at most their second member also satisfy this formula.

A sequence-member will satisfy this formula under two conditions: first if it is not an aardvark, secondly if it is flea-ridden. The only way to *fail* to satisfy it is by being both an aardvark and not flea-ridden. So Quine satisfies it, since he is not an aardvark. Moreover, if it is TRUE that all aardvarks are flea-ridden, then there is no object that we can put in the second position of a sequence that will fail to satisfy it. Consequently, if the formula is true, *all* sequences will satisfy it.

But what if the sentence is false, there is an aardvark that is not flea-ridden. Then no sequence will satisfy the sentence. For the relevant open sentence (i.e. the one formed by removing the universal quantifier) to be satisfied by a

particular sequence S , *all* sequences differing from S in at most one position must satisfy it. Once again, suppose we are concerned with the second position in our sequences (this is because we have symbolized our sentence using x''). Consider an arbitrary sequence S . If there is a non-flea-ridden aardvark, then at least one sequence will *not* be satisfied: that sequence with this aardvark in the second position (note: from the definition of "sequence," there will be a sequence with this object in this position, since for every object and every position, there is *some* sequence with that object in that position). Consequently, since for S to be satisfied ALL S' must be satisfied, S will not be satisfied: since either S itself will contain the non-flea-ridden aardvark, or some S' will.

So we see that satisfaction for universally quantified sentences is all-or-nothing: true ones are satisfied by all sequences, false ones by none.

We now have what we were seeking. For closed sentences, both quantified and nonquantified, satisfaction is all-or-nothing. Hence the definition of truth for closed sentences in terms of satisfaction in clause (h) makes sense: "a sentence is true if it is satisfied by all sequences." Given the way "satisfaction" has been defined, true sentences *will* be satisfied by all sequences, and false sentences by none.

Notes

Preface

- 1 *Tractatus Logico-Philosophicus*, 4.112. All the bibliographical references for Parts I and III are to be found at the back of the book.
- 2 The full story of the recent expansion of analytic movement is too complex to relate here, but some indications can be given. ESAP was founded in 1996, and has a significant Central European Section. National associations include: the Italian Society for Analytic Philosophy (SIFA) and its Romanian counterpart (SFRA), the Spanish Society for Analytic Philosophy (SEFA), the Portuguese Society for Analytic Philosophy (SPFA), Dutch-Flemish Association for Analytic Philosophy (VAF), the Croatian Society for Analytic Philosophy (CSAP), the French Society for Analytic Philosophy (SOPHA), and the German Society for Analytic Philosophy (GAP)—with the latter now being the second largest philosophical organization in Germany.
- 3 A success he did not live to see: in 1936 Schlick was gunned down on the steps of the University of Vienna by Johann Nelböck, a (mentally unstable) former student. At the ensuing trial, Nelböck claimed that Schlick's philosophical teachings had undermined his moral self-constraint—a line that was exploited by Austrian Nazis.

Chapter 1

- 1 All three were members of the Apostles, the covert and exclusive—and originally 12-membered—"Cambridge Conversazione Society." When Moore was invited to join in 1894, Russell and McTaggart were already members—in Russell's case since 1892. Of Moore's first appearance at a society meeting, Russell later reported to his then wife that Moore had "electrified" the Apostles, who previously had "never realized what fearless intellect pure and unadulterated really means" (Russell to Alys Pearsall Smith, February 18, 1894, Spadoni 1978, p. 24). Russell's future collaborator A. N. Whitehead became an Apostle in 1884; Keynes would be elected in 1903, and would later invite Wittgenstein to join the club.
- 2 Bradley was the first philosopher to be awarded the Order of Merit, in 1924, by King George V. He was also something of a recluse, and his colleague R. G. Collingwood notes in his *Autobiography* that "[A]lthough I lived within a few hundred yards of him for sixteen years, I never to my knowledge set eyes on him." Bradley harbored an intense dislike of cats, and on occasion took to shooting them at night in Merton College's grounds. McTaggart was not without his idiosyncrasies either, with Russell noting in his own autobiography that "McTaggart was even shyer than I was. I heard a knock on my door one day . . . a very gentle knock. I said 'come in' but nothing happened. I said 'come in' again louder. The door opened, and I saw McTaggart on

the mat. He was already president of the union, and about to become a fellow, and I was inspired and in awe on account of his metaphysical reputation, but he was too shy to come in, and I was too shy to ask him in. I cannot remember how many minutes this situation lasted, but somehow or other he was at last in the room" (Russell 1951, p. 88). McTaggart was fond of cats (unlike Bradley), and his preferred mode of transport around Cambridge was a tricycle.

- 3 Summarizing his philosophical methodology in "On Some Aspects of Truth" (1911) Bradley begins by "noticing some misunderstandings as to the method employed in ultimate enquiry by writers like myself. There is an idea that we start, consciously or unconsciously, with certain axioms, and from these reason downward. This idea to my mind is baseless. The method actually followed may be called in the main the procedure used by Hegel, that of a direct ideal experiment made on reality. What is assumed is that I have to satisfy my theoretical want, or, in other words, that I resolve to think. And it is assumed that, if my thought is satisfied with itself, I have, with this truth and reality. But as to what will satisfy I have of course no knowledge in advance. My object is to get before me what will content a certain felt need, but the way and the means are to be discovered only by trial and rejection. The method is clearly experimental."
- 4 This example is drawn from an early (pre-rebellion) essay of Russell's; the latter evidently had quite a profound appreciation of the phenomenological underpinnings of Bradley's system: "If I lie in a field on a hot day with my eyes shut, and feel sleepily the heat of the sun, the buzz of the flies, the slight tickling of a few blades of grass, it is possible to get into a frame of mind which seems to belong to a much earlier stage of evolution; at such times there is only what Bradley calls a 'vague mass of the felt'; I do not reflect on the outside causes of the various blurred and indistinct sensations, nor on the fact that I am feeling these sensations." (On the Distinction Between the Psychological and Metaphysical Points of View, 1894.)
- 5 See Sprigge 1993: Part II, ch.3 for more on this theme.
- 6 For a fuller treatment of this see Allard 2005.

Chapter 2

- 1 Feinberg and Kasrils 1969, p. 160.
- 2 Russell rated this paper of Moore's very highly; Gilbert Ryle characterized its influence thus: "'The Nature of Judgment' could be described as the *De Interpretatione* of early twentieth-century Cambridge logic" (Ambrose and Lazerowitz, 1970, p. 90).
- 3 In his various criticisms of Bradley's views of relations Russell sometimes misunderstands (or misrepresents) Bradley's true views—which, admittedly, are often difficult to discern, and developed over the years; see Candlish (2007, ch. 6) for a detailed exploration of a complex tale.
- 4 Some of Moore's other claims made in this period certainly seem to point us in this direction: "It seems plain that a truth differs in no respect from the reality to which it was supposed to correspond: e.g. the truth that I exist differs in no respect from the corresponding reality—my existence. So far, indeed, from truth being defined by reference to reality, reality can only be defined by reference to truth" (1901–2/1993, p. 21).
- 5 In Frege 1980, p. 169; Russell was here explaining why he could not accept Frege's theory of sense—for more on the latter see section 5.
- 6 For more on Bradley, Russell, and the issue of relations see Candlish 2007, Gaskin 2008, and Simons 2010; Simons remarks "The metaphysics of relations (unlike their logic) is still in its infancy" (p. 199).

Chapter 3

- 1 But Moore's views were well-received by some contemporary quarters, most notably by much of the Bloomsbury Group—whose members included Lytton Strachey, Virginia Woolf, Roger Fry, Vanessa and Clive Bell, E. M. Forster and John Maynard Keynes—who looked very favorably upon his revisionary and aesthetically oriented normative views. As Keynes observed in his "My Early Beliefs," they were selective in what they took from Moore: "What we got from Moore was by no means entirely what he offered us. He had one foot on the threshold of the new heaven, but the other foot in Sidgwick and the Benthamite calculus . . . We accepted Moore's religion, so to speak, and disregarded his morals . . . We set on one side, not only that part of Moore . . . which dealt with the obligation to act so as to produce by causal connection the most probable maximum of eventual good . . . but also the part which discussed the duty of the individual to obey general rules."
- 2 Moore's view was perhaps less radical than it initially appears from this passage alone, given that he also believed *conscious states* possessed far greater intrinsic value than anything else, as he states in the previous quotation. Nonetheless, a beautiful cosmos entirely lacking in conscious observers has some intrinsic value—more than its ugly counterpart—even if the same sort of cosmos *plus* conscious beings capable of experiencing its beauty has a great deal more.

Chapter 4

- 1 He would later describe this as the being "the most important event" in "the most important year" of his entire intellectual life (Schilpp 1944, p. 12). Russell relates the encounter with Peano thus: "I was impressed by the fact that, in every discussion, he showed more precision and more logical rigour than was shown by anybody else. I went to him and said, 'I wish to read all your works. Have you got copies with you?' He had, and I immediately read them all. It was they that gave the impetus to my own views on the principles of mathematics" (1959, p. 51).
- 2 A "propositional function" is an expression that consists of a predicate and a variable, which is converted into a proposition when a name is substituted for the variable: for example, if "Socrates" is combined with the propositional function "*x* is mortal" the result is the proposition "Socrates is mortal." Whereas the old logic treated "Socrates is mortal" and "All Greeks are mortal" as essentially similar, since both assign predicates to subjects, the new logic treats the general proposition very differently, symbolizing it as $(x)(Gx \rightarrow Mx)$, or "for all *x*, if *x* is Greek then *x* is mortal"—in short, as a conditional statement whose subject matter (or domain of quantification) is everything.
- 3 Logicians hold that all the concepts of mathematics can be defined in terms of logical concepts, and that all the truths of mathematics can be proved from the basic principles of logic. Although the term "logicism" has been in wide use since the 1930s, it was not used by the early logicians themselves. For more on the logicist project in general, see Stein 1998 and Proops 2006.
- 4 Peano's five primitive propositions are: (a) 0 is a number, (b) the successor of any number is a number, (c) no two numbers have the same successor, (d) 0 is not the successor of any number, and (e) if a set of numbers *S* contains zero and also the successor of every number in *S*, then every number is in *S*.
- 5 "The doctrine of types is here put forward tentatively, as affording a possible solution of the contradiction; but it requires, in all probability to be transformed into some subtler shape before it can answer all difficulties" (*Principles* Appendix B 1903, p. 523).

- 6 In the *Principles* Russell formulated the doctrine of types in terms of propositional functions and their “ranges of significance,” that is the kind(s) of objects a variable x can be if Fx is to constitute a genuine proposition.
- 7 In 1931 Gödel produced his “incompleteness theorems,” which established limitations on what a deductive system could prove. In brief, Gödel proved that a system that is powerful enough to express arithmetic will, if it is consistent, be incomplete, in the sense that there will be truths that are not deducible from the axioms. Although Gödel’s target was Hilbert’s formalist project, his results reduced the appeal of logicism, even if they did not undermine it directly. For more on the relationship between the incompleteness theorems and logicism, see Hellman 1981, and Leng’s contribution to this volume.

Chapter 5

- 1 For a fuller account of Frege’s treatment of number see <http://plato.stanford.edu/entries/frege/>
- 2 The appendix begins thus: “Hardly anything more unwelcome can befall a scientific writer than one of the foundations of his edifice be shaken after the work is finished. I have been placed in this position by a letter of Mr. Bertrand Russell just as the printing of the second volume was nearing completion. . . .”

Chapter 6

- 1 The heavily populated jungle of nonexistent, or merely “subsistent,” entities is entirely emptied. It is not unknown for accounts of “On Denoting” to present Russell as battling against Meinong, an influential proponent of the doctrine of nonexistent objects. In fact, since Russell had himself been an enthusiastic proponent of precisely the same doctrine a year or two earlier, it is just as accurate to view OD as a rejection of his own (earlier) view.
- 2 Here Russell rails against the notion that judgments involve, mental intermediaries: “The view seems to be that there is some mental existent which may be called the ‘idea’ of something outside the mind of the person who has the idea, and that since judgment is a mental event, its constituents must be constituents of the mind of the person judging. But in this view ideas become a veil between us and outside things—we never really, in knowledge, attain to the things we are supposed to be knowing about, but only to the ideas of those things. The relation of mind, idea and object, on this view, is utterly obscure, and, so far as I can see, nothing discoverable by inspection warrants the intrusion of the idea between the mind and object. I suspect that the view is fostered by the dislike of relations, and that it is felt the mind could not know objects unless there were something ‘in’ the mind which could be called the state of knowing the object” (1910–11, p. 119).
- 3 It would not be long before the class of genuine (or “logically proper”) names was shrunk further still: in the 1918 lectures “The Philosophy of Logical Atomism” the genuine names are further restricted to “this” and “that,” used to refer to sense-data.
- 4 Reprinted in *Mysticism and Logic and Other Essays* (1918); other relevant works in this period include *Our Knowledge of the External World* (1914), and “The Ultimate Constituents of Matter” (1915).
- 5 In fact he avoids this fate—and idealism itself—only by virtue of holding that sense-data are physical rather than mental in nature. Since we are *directly* aware of sense-data, if the latter are physical it is difficult to avoid concluding that they are identical with the neural processes in our brains that are triggered by inputs from

our perceptual systems—and Russell was tempted by this position; see Hylton 1990, pp. 361–75 for a good discussion of Russell’s views on this issue in this period.

Chapter 7

- 1 1931, MS 154, “The Cambridge Wittgenstein Archive,” www.wittgen-cam.ac.uk/
- 2 Dreben and Floyd 2011; the card dates from April 26, 1917.
- 3 The only book on philosophy, but not the only book: Wittgenstein also published a primer on spelling in 1926. The nonappearance of other philosophical works did not mean that Wittgenstein continued to subscribe to all the doctrines of the *Tractatus*—this was far from being the case, as we shall see in due course.
- 4 Wittgenstein’s alternative to the theory of types also relies on rules of syntax. Russell evades the class paradox by (in effect) introducing rules that restrict the ranges of objects predicates can apply to. Wittgenstein took a different path, and argued that in a properly constructed language the rules governing the possible ways functional expressions can combine with one another simply do not permit functions to apply to themselves, or (what amounts to the same thing) serve as their own arguments (TLP 5.251). That the relevant symbols are of different types is something that is *shown* by the symbols, it is not something that can be said.
- 5 One of the first to be puzzled was Frege: “What you write about the purpose of your book strikes me as strange. According to you, that purpose can only be achieved if others have already thought the thoughts expressed in it. The pleasure of reading your book can therefore no longer arise through the already known content, but, rather, only through the form, in which is revealed something of the individuality of the author. Thereby the book becomes an artistic rather than a scientific achievement; that which is said steps back behind how it is said” (Frege, letter to Wittgenstein September 16, 1919, Dreben and Floyd 2011, p. 57).
- 6 Wittgenstein himself made this plain in a letter to the publisher von Ficker: “[T]he book’s point is ethical. I once meant to include in the preface a sentence which is not in fact there now, but which I will write out for you here . . . my work consists of two parts: the one presented here plus all I have *not* written. And it is precisely this second part which is the important one. For the ethical gets its limit drawn from the inside, as it were, by my book; and I am convinced that this is the *ONLY rigorous* way of drawing that limit. In short, I believe that where *many* others today are just *gassing*, I have managed in my book to put everything firmly into place by being silent about it” Luckhardt 1979, p. 94.
- 7 For more on this, and good introductions to the *Tractatus* as a whole, see Kenny 2006 and Morris 2008.

Chapter 8

- 1 The two names can be taken as equivalent; the protagonists themselves were divided over which was the more appropriate—although by the mid-1930s “logical empiricism” was preferred by the leading members of the movement. The name “Vienna Circle” was a suggestion of Neurath’s, who wanted to exploit the pleasing connotations of the “Vienna woods” and the “Viennese waltz.” What is sometimes called the “First Vienna Circle” was a group comprising Philipp Frank, Hans Hahn, Otto Neurath, and Richard von Mises, who met from 1907–11 to discuss the implications of the revolutionary scientific work of Mach, Boltzmann, Planck, and Einstein; they sought to synthesize empiricism and the new mathematical and logical work of Poincaré, Hilbert, and Russell and Whitehead.

- 2 For additional and more in-depth coverage see Friedman 1999 and also Richardson and Uebel 2007.
- 3 For a comprehensive list of its “core” and “peripheral” members and participants see Stadler 2007, p. 15.
- 4 Neurath was one of the more politically active members: a leading Austro-Marxist, he served as economics minister in the Bavarian Soviet Republic in 1919.
- 5 Kant had argued that Euclidean geometry is hard-wired, as it were, into our minds, and hence that we could not conceive of space conforming to a different geometry. In fairness to Kant, the existence of consistent non-Euclidean geometries was not known to him.
- 6 Wittgenstein himself did not join the Circle, but did meet several times with Schlick, Carnap, and Waismann in 1927–9. Carnap later recalled: “Before the first meeting Schlick admonished us urgently not to start a discussion of the kind to which we were accustomed in Circle . . . We should even be cautious in asking questions, because Wittgenstein was very sensitive and easily disturbed by a direct question. The best approach, Schlick said, would be to let Wittgenstein talk, and then ask only very cautiously for the necessary elucidations. When I met Wittgenstein, I saw that Schlick’s warnings were fully justified . . . His point of view and his attitude toward people and problems, even theoretical problems, were much more similar to those of a creative artist than to those of a scientist; one might almost say, similar to those of a religious prophet or a seer” (Schlipp 1963, p. 25).
- 7 The “Manifesto” in full can be found in Neurath’s *Empiricism and Sociology* (1973, p. 301–18).
- 8 The passage continues in a quasi-poetic manner: “. . . The content, soul and spirit of science is lodged naturally in what in the last analysis its statements actually mean; the philosophical activity of giving meaning is therefore the Alpha and Omega of all scientific knowledge” (1959, p. 56).
- 9 “What is Metaphysics?” was Heidegger’s inaugural address to the faculties of the University of Freiburg on July 24, 1929 (see Krell 1993, pp. 93–110 for an English translation). Carnap drily remarks “We could just as well have selected passages from any other of the numerous metaphysicians of the present or of the past; yet the selected passages seem to us to illustrate our thesis especially well.” Indeed. For an illuminating take on the relationship between Carnap and Heidegger, and their encounter at Davos in 1929, see Friedman 2000.
- 10 Ayer goes on to say: “I don’t think this condemns the principle, since I think that a principle may be intuitively clear and effective even though you don’t get a watertight formal statement of it. But of course it would be much nicer to get a watertight formal statement of it, if this were possible.” —“Conversation with A.J. Ayer” in Magee 1971. For a useful discussion of the verification principle in general, and Ayer in particular, see Foster 1985.
- 11 Where Stevenson writes: “. . . the present work finds much more to defend in the analyses of Carnap, Ayer and the others, than it finds to attack. It seeks only to qualify their views . . . It hopes to make clear that ‘emotive’ need not itself have a derogatory emotive meaning. And in particular, it emphasizes the complex descriptive meaning that ethical judgments can have, in addition to their emotive meaning” (1944, p. 267).
- 12 Blackburn 1984, 1998; Gibbard 1990.
- 13 For an overview and discussion of Carl Hempel’s work see <http://plato.stanford.edu/entries/hempel/>. For Ernst Nagel’s work on intertheoretic reduction and intertheoretic relations, see <http://plato.stanford.edu/entries/physics-interrelate/>
- 14 Popper’s antipathy toward Wittgenstein came to a head in a notorious (the details are much disputed) confrontation when Popper gave a paper to the Moral Sciences Club in Cambridge in 1946; for a readable take on this encounter see Edmonds and Edinow 2001.

- 15 Looking back on his relationship with Popper in their Vienna days, Carnap later wrote: "Among philosophers in Vienna who did not belong to the Circle I found the contact with Karl Popper most stimulating . . . I remember with pleasure the talks with him and Feigl in the summer of 1932, in the Tyrolean Alps. His basic philosophical attitude was quite similar to that of the Circle. However, he had a tendency to overemphasize our differences. In his book he was critical of the 'positivists', by which he seemed to mean chiefly the Vienna Circle . . . He thereby antagonized some of the leading figures in our movement, e.g. Schlick, Neurath, and Reichenbach. Feigl and I tried in vain to effect a better mutual understanding and philosophical reconciliation" (Schlipp 1991, p. 31).
- 16 Not least because of his impact and influence on Quine, whom we will be looking at in due course.
- 17 A tactic anticipated (informally) by Russell in his introduction to the *Tractatus*: "These difficulties suggest to my mind some such possibility as this: that every language has, as Mr Wittgenstein says, a structure which, *in the language*, nothing can be said, but that there may be another language dealing with the structure of the first language, and having itself a new structure, and that to this hierarchy of languages there may be no limit. . . . Such an hypothesis is very difficult, and I can see objections to it which at the moment I do not know how to answer. Yet I do not see how any easier hypothesis can escape from Mr Wittgenstein's conclusions" (*TLP*, p. xxii).
- 18 As a letter from Neurath to Carnap in October 1932 forcefully reveals: "The tear is running . . . It'd be to throw up, if one didn't have to laugh. And behind it all stands Hitler . . . Here comes God and Religion to the front and ancestral truths and the German *Volk*, and what you need to stab a jewish socialist with knife between the ribs . . . Oh Carnap! Oh World!" (Galison 1990, p. 742).
- 19 They were by no means alone in believing this: the Bauhaus movement had similar aspirations. Galison summarizes the similarities between the philosophical and artistic movements thus: "Both enterprises sought to instantiate a modernism emphasizing what I will call 'transparent construction', a manifest building up from simple elements to all higher forms that would, by virtue of the systematic constructional program itself, guarantee the exclusion of the decorative, mystical or metaphysical" (1990, p. 710). Wittgenstein bridged the two, working closely on the design of the Stonborough house in Vienna, 1926–8. One of Wittgenstein's sisters would characterize it as "hausgewordene Logik" (logic become house).

Chapter 9

- 1 Ramsey was a precocious talent whose career was cut short: he died in 1930 aged 29.
- 2 Russell and Moore were the examiners; the latter commented in his examiners report: "It is my personal opinion that Mr Wittgenstein's thesis is a work of genius; but, be that as it may, it is certainly well up to the standard required for the Cambridge degree of Doctor of Philosophy." For an amusing—and informative—take on the *viva*, and a different assessment of the work's merits, see Goldstein 1999.
- 3 Report to the Council of Trinity on Wittgenstein's work (Russell 2009, p. 420).
- 4 One of these problems concerned the Tractarian doctrine that all elementary propositions are logically independent of one another. Wittgenstein came to see that the internal relations between sensible properties—if an object is red all over at a time *t* it cannot also be blue or yellow all over at *t*—had to be accommodated, and sketched one possible approach in his 1929 "Some Remarks on Logical Form"—the only philosophical paper Wittgenstein would publish during his lifetime.

- 5 This enterprise was assisted (and made more difficult) by the subsequent publication of many of Wittgenstein's notebooks and manuscripts, for example, *Remarks on the Foundations of Mathematics* (1956), *The Blue and Brown Books* (1958), *Philosophical Remarks* (1964), *Lectures and Conversations on Aesthetics, Psychology and Religious Belief* (1966), *Zettel* (1967), *On Certainty* (1969), *Philosophical Grammar* (1974), *Culture and Value* (1980), *Remarks on the Philosophy of Psychology* (1980), *Last Writings on the Philosophy of Psychology* (vol. 1 1982, vol. 2 1992), and *The Big Typescript* (2005). For a more complete list see <http://plato.stanford.edu/entries/wittgenstein/>
- 6 "A commonsense person, when he reads earlier philosophers thinks—quite rightly—'Sheer nonsense.' When he listens to me, he thinks—rightly again—'Nothing but stale truisms.' That is how the image of philosophy has changed" (MS 219, p. 6, quoted in Kenny 1984, p. 57).
- 7 Indeed, on the same page Russell tells us "It is not an altogether pleasant experience to find oneself regarded as antiquated after having been, for a time, in the fashion. It is difficult to accept this experience gracefully."

Chapter 10

- 1 Hylton also observes that because Quine opted to concentrate on mathematics rather than philosophy in his student days, "his formal study of philosophy, as distinct from logic, was not extensive. As an undergraduate he took two survey courses in the subject, which evidently made little impression. . . . Quine thus began his career with little background in philosophy. He does not seem to have done much systematic work to fill the gaps. With the notable exceptions of Russell, and especially Carnap, he is not, in his early work, either reacting against or building upon the work of others" (2007, p. 33).
- 2 For an interesting and recent reappraisal of Quine's criticisms of Carnap, see Gary Ebbs 2011.
- 3 Or as Quine himself puts it: "Carnap, Lewis and others take a pragmatic stand on the question of choosing between language forms, scientific frameworks; but their pragmatism leaves off at the imagined boundary between the analytic and the synthetic. In repudiating such a boundary I espouse a more thorough pragmatism" (1953, p. 46).
- 4 *Beacon Hill Paper*, May 15, 1996, p. 11; www.wvquine.org/wvq-newspaper.html
- 5 Quine expands thus: "Translation between kindred languages, e.g. Frisian and English, is aided by resemblance of cognate word forms. Translation between unrelated languages, e.g., Hungarian and English, may be aided by traditional equations that have evolved in step with a shared culture. What is relevant to our purpose is *radical* translation, i.e., translation of the language of a hitherto untouched people. The task is one that is not in practice undertaken in its extreme form, since a chain of interpreters of a sort can be recruited of marginal persons across the darkest archipelago . . . I shall imagine that all help of interpreters is included" (1960, p. 28).
- 6 The "complement" of a rabbit is the entirety of the universe *apart from the rabbit*; a "rabbit-slice" is short for "temporal slice of a rabbit," that is, a single brief phase of the rabbit construed as a four-dimensional object—one that extends through time as well as space, and has temporal parts as well as spatial parts.

Chapter 11

- 1 Here is Ryle recalling the depth and manner of Wittgenstein's influence: "Philosophers who never met him—and few of us did meet him—can be heard talking philosophy in his tones of voice; and students who can barely spell his name now wrinkle up their noses at things which had a bad smell for him" (1971, p. 249).
- 2 As Warnock reminds us, whether differences of approach or opinion are major or minor is, to an extent, context-relative: "It can only have been from a *very* great distance, or through glasses of highly imperfect focus, that everyone [in Oxford] at that time looked much alike, like devotees of a 'school'" (1976, p. 51).
- 3 Hacker here nicely captures the unusual character of these meetings: "Among those who attended his 'Saturday Mornings' over the years were Marcus Dick, Grice, Hampshire, Hare, Hart, P. H. Nowell-Smith, Paul, Pears, Strawson, Urmson, Warnock and A. D. Woozley. The subjects discussed ranged far and wide. . . . The general topics handled were equally diverse, including a term spent discussing rules of games (with an eye to questions about meaning and rules for the use of words), as preparation for which each member of the group was given a book of rules to study; and aesthetics, for which an illustrated handbook of industrial design containing pronouncements on the design of humble artefacts was scrutinized in order to find out what people *actually* say in aesthetic appraisal when the topic is not too grand to inhibit good sense Apropos Wittgenstein's comparison of words to tools, the expressions 'tool,' 'instrument,' 'implement,' 'utensil,' 'appliance,' 'equipment,' 'apparatus,' 'gear,' 'kit,' 'device' and 'gimmick' were examined in patient detail (were kitchen scissors, garden shears, dress-making scissors, surgeon's scissors utensils, tools or implements?) with a view to determining the most helpful analogy" (1996c, p. 151).
- 4 In a note at the end of the article Ryle explains why he has avoided Russell's terminology: "In this paper I have deliberately refrained from describing expressions as 'incomplete symbols' [as Russell did] or quasi-things as 'logical constructions.' Partly I have abstained because I am fairly ignorant of the doctrines in which these are technical terms, though in so far as I do understand them, I think that I could re-state them in words which I like better without modifying the doctrines. But partly, also, I think that the terms themselves are rather ill-chosen and apt to cause unnecessary perplexities. But I do think that I have been talking about what is talked about by those who use these terms, when they use them" (op.cit., p. 170).
- 5 The use of "presupposition" here is something of an anachronism since Strawson himself would only use the term in this sort of context in his 1964 'Identifying Reference and Truth-Values,' but it has subsequently entered the philosophical lexicon. For a useful survey of recent debates between Russellians and Strawsonians see Ramachandran 2008. For more on subsequent work on presupposition, including the issue of whether presupposition is semantic (something sentences have) or pragmatic (something sentence-using people have) see Barry Smith's contribution to this volume, and <http://plato.stanford.edu/entries/presupposition/#FreStrTra>
- 6 To illustrate, suppose, for example, the master sound includes five tones, C-D-E-F-G, of distinct but differing pitches, and that every time one passes D one also hears the first bars of Beethoven's Fifth Symphony, and that every time one passes F one hears the Beatles' "Yesterday"—the notion that these two tunes are *located* at D and F, and exist at these master sound locations even when you are elsewhere, now seems quite plausible.
- 7 "Conversation with Peter Strawson," Magee 1971, p. 124.
- 8 Grice himself left Oxford for Berkeley in 1967—he had been at Oxford since 1938. The ordinary language approach no longer dominates in the way it once

did, but neither has it completely vanished; for a recent exemplar of the genre, see Dennett's "Sakes and dints" (2012), available at <http://ase.tufts.edu/cogstud/papers/TLS2012.pdf>

Chapter 12

- 1 Occasionally her absolutist views show through: "... if someone really thinks *in advance*, that it is open to question whether such an action as procuring the judicial execution of the innocent should be quite excluded consideration—I do not want to argue with him; he shows a corrupt mind" (1958, p. 15).
- 2 There is a further respect in which Anscombe's essay was prophetic: the "discourse ethics" promulgated by Apel (1988) and Habermas (1998) is precisely a way of deriving ethical principles from the conventions that make communicative language possible. In *Moral Luck* (1981) and *Ethics and the Limits of Philosophy* (1985) Bernard Williams argued—in a manner at times echoing Anscombe—that much of received morality is redundant in a secular age: the "various features of the moral system support each other, and collectively they are modelled on the prerogatives of a Pelagian God. The strictness of the criteria for judgment responds to the supposed immensity of what is handed out . . . they collectively invite the skepticism I have mentioned. They face a problem of how people's character or dispositions could ever be the object of such a judgment. They are unlikely to be fully responsible for them, and it is even less likely that we can know to what extent they are responsible for them" (1985, p. 38). In a different vein, but roughly the same period, Mackie's *Ethics Inventing Right and Wrong* (1977) was an influential challenge to ethical realism.
- 3 For more on Rawls and his theory of justice see <http://plato.stanford.edu/entries/rawls/>
- 4 Rawl's method of reflective equilibrium consists of working back and forth between judgments (or intuitions) about particular cases, and the general moral principles that govern them, modifying both where necessary, with the aim of achieving a maximally coherent system.

Chapter 13

- 1 Other relevant papers from this period can be found in Davidson's *Inquiries into Truth and Interpretation* (1984; 2nd edn 2001).
- 2 For more on Tarski's theory of truth see the Appendix to Part I.
- 3 For one assessment see Soames 2003b, vol. 2, chs 12–13. Although critical of Davidson, Soames is by no means dismissive of his achievement: "Whatever the shortcomings in conception and execution, his overall truth-theoretic approach to meaning in natural language represented a major advance over both the barren semantic skepticism of Quine, and the anti-theoretical, yet philosophically overreaching, linguistic methodology of Wittgenstein and the ordinary language philosophers. For more see Hahn 1999, LePore and Ludwig 2005, 2007, and Amoretti and Vassalo 2009.
- 4 Davidson concludes "On the Very Idea of a Conceptual Scheme" with these words: "In giving up the dualism of scheme and world, we do not give up the world, but reestablish unmediated touch with the familiar objects whose antics make our sentences and opinions true or false." In his later "The Structure and Content of Truth" he writes: "We should not say that truth is correspondence, coherence, warranted assertability, ideally justified assertability, what is accepted in the conversation of the

right people, what science will end up maintaining, what explains the convergence of single theories in science, or the success of our ordinary beliefs. To the extent that realism and anti-realism depend on one or another of these views of truth we should refuse to endorse either. Realism, with its insistence on radically nonepistemic correspondence, asks more of truth than we can understand; antirealism, with its limitation of truth to what can be ascertained, deprives truth of its role as an intersubjective standard" (1990, p. 309).

- 5 For one line of argument leading to the conclusion that Davidson is close to being an idealist see Nagel 1986, pp. 90–109.

Chapter 14

- 1 Both philosophers have contributed to many other areas of philosophy—both are logicians, for example—and the doctrines we will be focusing on here represent a small part of their overall output. See de Gaynesford 2006 for an overview of Putnam's work, and Burgess (forthcoming) and Berger 2011 for more on Kripke.
- 2 The "new" (or "direct") theory of reference that Kripke and Putnam are both associated with was also advocated—and influenced by—a number of others, for example, Keith Donnellan, David Kaplan, Ruth Barcan Marcus, and John Perry.
- 3 In Soames' view it rates "among the most important works ever [in the philosophy of language], ranking with the classical work of Frege in the late nineteenth century, and of Russell and Tarski in the first half of the twentieth" (2003, vol. II, p. 336). The influence of Kripke's work was felt before the publication of *Naming and Necessity* in 1980: the Princeton lectures were taped, and a transcript of them by Gilbert Harman and Thomas Nagel was circulated.
- 4 One (somewhat) promising option is a minimalist "meta-linguistic" approach, which ascribes to a proper name NN a descriptive sense along the lines of "the bearer of name 'NN.'" For further details see Bach 1987 and Sainsbury 2005; for a useful overview: <http://plato.stanford.edu/entries/names/>
- 5 Attempts to solve these problems in the direct reference tradition include Salmon 1986 (2nd edn), Perry 2001, and Soames 2002.
- 6 Quine frames one of his complaints thus: "Perhaps I can evoke the appropriate sense of bewilderment as follows. Mathematicians may conceivably be said to be necessarily rational and not necessarily two-legged; and cyclists necessarily two-legged and not necessarily rational. But what of an individual who counts among his eccentricities both mathematics and cycling? Is this concrete individual necessarily rational and contingently two-legged or vice-versa? Just insofar as we are talking referentially of the object, with no special bias towards a background grouping of mathematicians against cyclists or vice versa, there is no semblance of sense in rating some of his attributes as necessary and others as contingent. Some of his attributes count as important and others as contingent, yes, some as enduring and others as fleeting; but none as necessary or contingent" (1960, p. 199).
- 7 The subtitle of the anthology is "Twenty Years of Reflection on Hilary Putnam's 'The Meaning of 'Meaning.''"
- 8 Another important line of argument pointing in the same general direction, involves a different kind of environmental fact: differences in the *beliefs* of other people in one's linguistic community can (it is argued) go to determine the contents of one's own beliefs. This was anticipated (in his discussion of "elms") by Putnam in "The Meaning of 'Meaning,'" and developed in great depth in Burge 1979.

- 9 For further reading and background see the following Stanford Encyclopedia entries: “Narrow Mental Content,” “Externalism About Mental Content,” and “Singular Propositions.”

Chapter 15

- 1 For an overview of the various forms of functionalism, and some of the debates to which the doctrine has led in recent years, see <http://plato.stanford.edu/entries/functionalism/>
- 2 For classic statements of Fodor’s position see his *Language of Thought* (1975), and *Psychosemantics* (1987).
- 3 There have always been a (small) minority of analytic philosophers who espoused dualism—or even idealism—and this remained the case in the period in question, for example, Robinson 1982 and Foster 1983, 1991. Nagel, Searle, McGinn, and Strawson all rejected dualism: it was an improved version of *physicalism* that they were calling for.
- 4 The idea that the relationship between consciousness and the brain was a uniquely difficult problem, and that little or no progress had been made on solving it was very much in the air at the time, and not only among philosophers. In *The Astonishing Hypothesis* (1994) Francis Crick forcefully argued that consciousness should now be viewed as a proper object of scientific investigation. Neuroscientists such as Damasio (1999) and Edelman (1992) had already concluded as much—and even physicists were having their say, for example, Penrose (in the *Emperors New Mind*, 1989). In 1994 the first Tucson “Towards a Science of Consciousness” conference took place, which led to the founding of the *Association for the Scientific Study of Consciousness*. The multi-disciplinary *Journal of Consciousness Studies*, was founded in the same year.
- 5 Given that our world is the actual world, and that in our world water is H_2O , and that “water” is a rigid designator (i.e. it refers to the same thing in all other possible worlds as it does in our world), then “water is H_2O ” is true in all worlds, and so is a necessary truth. But according to Chalmers, there is another sense in which water is *not* necessarily H_2O . There are worlds that resemble our world in superficial respects—where there is watery stuff that flows from taps, falls as rain, fills rivers, etc.—but where the chemical constitution of this stuff is X_YZ , and not H_2O . Since the inhabitants of this world call their watery stuff “water,” in this world water is X_YZ and not H_2O . Indeed, if this world were the actual world, then “water is X_YZ ” would be a necessary truth, since it would be true in all other possible (i.e. counterfactual) worlds. So in thinking about modality we need to think in terms of two dimensions: the possible worlds are still conceived as alternative ways the universe might be, but we also have to take into account the consequences of differing possible worlds being actual. What the inhabitants of the H_2O and X_YZ worlds *mean* by “water” also has two aspects or dimensions. There is a narrow aspect, a “primary intension” in Chalmers’ terms, which is (roughly) the general conception of *water-like stuff*, which the inhabitants of both worlds share. There is also the “secondary intension,” the (differing) sorts of stuff picked out by the primary intension in different worlds. Chalmers’ construes primary intensions in epistemic terms: they are the component of meaning to which we have complete a priori access. In the case of terms such as “water,” what a given primary intension picks out in a given world will depend on scientific investigation, and so is a posteriori. It is the primary intensions that fix the limits of the conceivable for a given term or concept. Within this framework, or so Chalmers argues, there is a sense in which what is conceivable *is* logically possible.

Furthermore, since zombies fall into this category, we should conclude that physicalism, at least of the reductive variety, is false: "... the primary intension determines a perfectly good property of objects in possible worlds. The property of being watery stuff is a perfectly reasonable property, even though it is not the same as the property of being H₂O. If we can show that there are possible worlds that are physically identical to ours but in which the property introduced by the primary intension is lacking, then dualism will follow. This is just what has been done with consciousness. We have seen that there are worlds just like ours that lack consciousness, according to the primary intension thereof. This difference in worlds is sufficient to show that there are properties of our world over and above the physical properties" (1996, pp. 132–3).

- 6 His position on this issue remained unchanged for many years: in the section of *My Philosophical Development* entitled "My Present View of the World" Russell tells us that he maintains "an opinion which all other philosophers find shocking; namely, that people's thoughts are in their heads. The light from a star travels over intervening space and causes a disturbance in the optic nerve ending in an occurrence in the brain. What I maintain is that the occurrence in the brain is a visual sensation. I maintain, in fact, that the brain consists of thoughts—using 'thought' in its widest sense, as it is used by Descartes . . . What I maintain is that we *can* witness or observe what goes on in our heads, and that we cannot witness or observe anything else at all" (Russell 1959, pp. 18–19).

Appendix

- 1 What follows draws heavily on the expositions in Blackburn's *Spreading the Word* (1984) and Platts' *Ways of Meaning* (1979); for a more comprehensive overview, see <http://plato.stanford.edu/entries/tarski-truth/> and <http://plato.stanford.edu/entries/tarski/#Tru>

Part II

Current Research and Issues

Introduction to Part II

Barry Dainton and Howard Robinson

By this point we have covered the historical and formal background to contemporary analytical philosophy. From here the baton is taken up by the contributors, who will develop what they take to be the salient points in the various areas of the discipline.

The first five chapters deal with what one might call the “hard core” analytical issues: the philosophy of mathematics, the philosophy of language, the philosophy of science in general, and the philosophy of physics in particular.

Mary Leng takes us a good way into the philosophy of mathematics, providing en route enough of the relevant background to make recent debates intelligible. Although the philosophy of mathematics is generally regarded as a specialized and technical branch of philosophy, even by analytical philosophers, it was from the determination of thinkers such as Frege and Russell to find a solid foundation for mathematics that analytical philosophy developed in the way it did, as we saw in Part I. As we have also seen, Frege’s attempt to ground mathematics in logic was undermined by Russell’s paradox. Starting from Hume and Kant, Leng guides us through nineteenth-century developments, and reactions in the twentieth century to the problems encountered by Frege’s logicist project; these include Brouwer’s intuitionism, Hilbert’s attempts to ground infinitary mathematics in finite consistency proofs, and the undermining of Hilbert’s program by Gödel. In turn, Gödel’s Platonism is challenged by the epistemological problems raised by Benacerraf. Leng also discusses the problems faced by Quine’s empiricist approach to mathematics, Field’s fictionalism, the neologicism of Wright and Hale, and Shapiro’s structuralism, among others. All in all, if the attempt to demystify mathematics seems no nearer to a generally accepted conclusion, the range of answers to the question “What is mathematics?” is broader than it was, and the strengths and weaknesses of the competing positions are better understood.

One of the main and most obvious lessons of Part I is that the history of analytic philosophy has to a large extent been a history of analytic philosophers engaging with the philosophical issues thrown up by language. While there is a great deal more to analytic philosophy than the philosophy of language, it is also true that language has long been a central concern of the analytic school,

and we have two essays devoted to this topic. Barry Smith's "Philosophy and Language" introduces us to some of the major issues that have engaged analytic philosophers in more recent years.

It is thanks to our mastery of language that we can communicate as well as we do, but communication comes in many forms, not all of which are linguistic, for example, much can be communicated by a nod of the head, or the wink of an eye. What is the defining or distinctive feature of *linguistic* communication? This sort of communication clearly relies on language, but this merely brings to the fore a further question: what is a *language*? For Quine the answer is straightforward: a language is a set of well-formed sentences. But as Chomsky influentially argued, this Quinean account may be simple, but it is also problematic. For Chomsky, a language exists primarily inside the minds of its speakers, in the form of a system of grammatical rules that are not accessible to consciousness. As for linguistic communication itself, there is a straightforward account (Smith dubs it "naïve") of what it involves: sentences have standard meanings, and this meaning fully determines what a speaker can assert on any given occasion when using a sentence. However, this naïve picture does not sit easily with the fact that we often do not mean (or convey to our audience) what we literally say. The standard way of addressing this issue is to distinguish between *semantics* and *pragmatics*. In this division of labor the former discipline deals with the stable literal meaning of sentences, and the latter deals with how speakers frame sentences to mean what they do in particular contexts of utterance. This distinction is in play in Grice's work: in Part I (Chapter 11) we saw how we distinguish between the standard stable meanings of sentences and what these sentences *implicate* when used on particular occasions. Quite how to draw the line between semantics and pragmatics has been a vexed issue in recent work. Whereas some have been content merely with comparatively mild enhancements to the Gricean picture, others have advocated the overthrow of semantics by pragmatics: for these radicals, there are no stable context-independent meanings for semanticists to deal with. Smith guides us through these competing positions, and the often subtle linguistic considerations driving them.

If there is a single fundamental issue in the philosophy of language it is simply this: What is the nature of meaning? This is the issue Richard Gaskin focuses on in his "Meaning, Normativity, and Naturalism." In 1982 Kripke published *Wittgenstein on Rules and Private Language: An Elementary Exposition*. This short book made an immediate impact on the analytical scene. In it Kripke offered a radically original interpretation of Wittgenstein's *Philosophical Investigations*. Wittgenstein's so-called private language argument had long been one of the most discussed aspects of the entire *Investigations*. The upshot of the argument is that it is incoherent to suppose there could be a language that by its very nature could be understood by only one person. (Since our ordinary

expressions for sensations would have private meanings of just this kind on some conceptions of the mind—or so it has often been argued—the ramifications of argument are potentially considerable.) The private language argument was usually taken to start at §242 of the *Investigations*, but Kripke argues that this is a mistake, and that Wittgenstein's principal argument against private language has already been completed. At §201 Wittgenstein tell us "This was our paradox: no course of action could be determined by a rule, because every course of action can be made out to accord with a rule." According to Kripke, the paradox in question is "the most radical and original skeptical problem that philosophy has seen to date." As developed by Kripke, this paradox threatens the very notion there can be a fact of the matter as to what any of us mean by any of our words. In his contribution Gaskin expounds and assesses Kripke's interpretation. The true import of the paradox, or so Gaskin argues, is not that there is no fact of the matter about what any of us mean, but that meaning is irreducible: "In grasping a rule [for the correct use of a word] we—individually and collectively—are put in touch with the ideal, in a way that goes beyond the merely causal."

The close relation of science to philosophy was especially important to the logical positivists. This importance took two forms. One was a concern with the nature of scientific explanation in general, the other was the more specific task of trying to make sense of the achievements of early twentieth-century physics in the shape of the revolutionary theories of relativity and quantum theory. James Ladyman deals with science in general and Barry Loewer with physics in particular.

Ladyman's first main topic is scientific methodology. If science is trusted in contemporary society in a way that other disciplines are not—for example, astrology, or voodoo—it is, presumably, because science has a distinctive methodology. But what is this scientific methodology, and why has it proved so uniquely successful? Perhaps surprisingly, given the success of science, there is little consensus on this topic. Philosophers of science have proposed accounts of scientific methodology that are strikingly (perhaps worryingly) different. After surveying the more influential of these accounts, Ladyman moves on to the *metaphysics* of science. Many branches of science aim to discover the laws of nature, but what are natural laws? Again, there are very different accounts on offer, accounts that have significant implications for the picture of the world painted by science. Ladyman then moves on to scientific epistemology, and to the debate between realists, those who consider it reasonable for us to believe that science is closing in on the truth, and the anti-realists who deny this. The chapter closes with a look at the philosophy of specific sciences—such as chemistry, biology, psychology, cognitive science, computer science—all of which have undergone rapid development in recent years.

The main job of physics, as Barry Loewer notes, has always been “accounting for the motions of quotidian material objects.” Although important work on this front was done in ancient times it was not until the seventeenth century that physics emerged as an exact mathematical science, in the form of Newton’s mechanics, as elaborated in his *Principia Mathematica* (1687). In setting out the essentials of Newton’s physics, Loewer brings into clear focus those aspects of the system that would later prove problematic. The temporal neutrality of its laws is one such. How this neutrality (or symmetry) can be reconciled with the temporally asymmetric second law of thermodynamics became an increasingly pressing problem during the nineteenth century. Boltzmann’s statistical particle mechanics was an important breakthrough, but as Loewer explains, it also introduced problems whose ramifications are still being explored by physicists and philosophers. By the end of the nineteenth century a further tension in physics had emerged: between classical mechanics and Maxwell’s electromagnetic theory. In 1905 Einstein resolved the tension in an elegant way with his special theory of relativity. According to the latter there is no absolute distinction between time and space, and Newton’s doctrine that time flows at an equal rate through all of space is false. The metaphysical implications of the Einsteinian revolution continue to be debated by philosophers, as do the ramifications of quantum theory. This was devised as a response to the discoveries in particle physics made in the early decades of the twentieth century. But while quantum theory is brilliantly successful in accommodating the known behavior of atoms and subatomic particles, it also gives rise to new problems. Not the least of these can be stated simply: it is quite unclear what the physical world *could be like* if quantum theory provides an accurate and complete description of it. In guiding us through the differing responses to these quantum conundrums Loewer also vividly illustrates the very distinctive way in which analytic philosophers are engaged with developments in fundamental science.

If science, as a human activity, is concerned with *explanation*, the world itself is, one might think, driven by *causation*. This view was challenged by Russell, who thought that causation played no role in the fundamental engine of the universe. But the notion of causation nevertheless continues to play a central role in philosophical thinking, and within the analytic tradition Hume’s account of it has conditioned the debate. According to Hume, although we are naturally inclined to think that causes *make* their effects happen, this view does not stand up to close scrutiny; causation-in-the-world does not involve any form of necessitation, it is simply a matter of regularities. As with many Humean positions—on the existence of the self and induction, for instance—although philosophers find it hard to accept his position, there is a compelling quality about his argument that they find hard to ignore. Hume presents the great challenge: “if you think there is more to it than this, say clearly what that

‘more’ is!” It invariably proves very hard to meet this challenge. Helen Beebe looks at contemporary responses to Hume’s stance on causation. Simple forms of neo-Humean regularity theory face severe problems; David Lewis’ counterfactual analysis of causation is an influential attempt to overcome these problems, but it too is problematic on several fronts. Other contemporary philosophers—such as Armstrong—are prepared simply to deny Hume’s claim that we cannot make sense of necessary connections among events. After discussing these approaches, Beebe introduces us to some others: the “process” theories of Salmon and Dowe, probabilistic conceptions of causation, and the accounts developed in the recent causal modeling literature. In completing her wide-ranging survey she looks at recent debates concerning what the causal relata are (events, facts, or objects), the different solutions to the problem of isolating the causally relevant features of events, and whether there really is a univocal notion of cause in the first place.

As we saw in Part I—and will see again in our survey of recent developments in Part III—the role metaphysics plays within the analytic tradition has been a particularly contested one. The logical positivists were particularly hostile to it, arguing that metaphysical claims were cognitively meaningless. While they did not go quite so far, those working within ordinary language movement had little time (or inclination) to engage in grandiose metaphysical theorizing; they were more likely to think that metaphysical problems were pseudo-problems, to be dissolved rather than solved. But as we have also seen, since the 1970s, and the work of Putnam and Kripke, analytic philosophers has seen something of a “metaphysical turn.” The willingness to engage directly with traditional metaphysical issues is well exemplified by Jonathan Lowe’s contribution. Lowe focuses on three central metaphysical topics: identity, change, and modality. These are very basic concepts, but they are also related. When an object undergoes change—for example, if your cat loses a hair—it undergoes an alteration of its properties, but in another sense it remains the same: it is still *your cat* that is involved throughout. It is also the case that there are only certain changes that it is *possible* for an object to undergo; a cat can lose a hair, but it cannot be squeezed into the shape of a cube, unlike a lump of clay, which can. Quite how we should make sense of these familiar phenomena has troubled philosophers since ancient times, and a number of paradoxes and puzzles have led some metaphysicians to doubt whether it is even possible to develop a coherent account of change. In taking us through some of these paradoxes, and providing solutions to them, Lowe introduces us to his own views of identity, change, and modality, and the relationship of the latter to the concept of *essence*. Since Lowe also discusses how, on his view, metaphysical issues are distinct from empirical ones, the chapter’s aims are *meta*-metaphysical as well as purely metaphysical.

As our contributors show, the original analytic concerns with logic, language, and science are very much alive today, but analytic thought is by no means confined to these issues. It might be true to say that the human and practical now rival them as analytic concerns. The issues we have chosen to represent in this area are, on the human theoretical side, the ever-perennial mind-body problem, the theory of personal identity, and the epistemologically central issue of perception: and among the more practical, ethics and political philosophy.

As we saw in Part I (Chapter 15) the arguments of Nagel, Jackson, and Chalmers are meant to show that knowing all the physical facts still leaves something unknown, namely the subjective, *what it is like* or the qualitative nature (qualia) of experience. Hence the physical facts are not all the facts and physicalism is false. Daniel Stoljar's chapter takes up the argument from Jackson's famous thought experiment concerning a neuroscientist confined to a black and white room, and what she can (and cannot) deduce about the nature of color experience prior to having any. He starts from Jackson's observation that standard physicalism rests on an extremely optimistic view of our cognitive powers, and develops the view that, if we take into account the limitations that follow from our evolutionary situation, our failure to solve the mind-body problem becomes less threatening. It is an assumption of physicalism—again in its standard form—that the physical world can be completely characterized by concepts of the kind we currently possess. If we reject this assumption—and there are powerful reasons for so doing—then a more realistic (if modest) form of physicalism turns out not threatened in the least by any of the familiar qualia-based anti-physicalist arguments. Stoljar then moves on to consider whether this “epistemic” solution to the problem of consciousness can be extended to other longstanding problems in the philosophy of mind, such as the nature of intentionality and self-knowledge. Little work has yet been done on this question, but Stoljar goes on to suggest that useful light can be shed on it by distinguishing *descriptive* from *foundational* problems in psychology.

The mind-body issue is very closely connected to the question of what it is to be a person, and this, in turn, seems most often to be discussed via the problem of personal identity. Mark Johnston believes that we are naturally committed to the view that persons are basic or primitive entities. The alternative is that we are, as he puts it, “ontological trash,” which (roughly) means that our individuation is a matter of interest or perspective, rather in the way in which one's answer to the puzzle of Theseus' ship depends on whether one is a working sailor or an antiquarian. The various thought experiments in the literature that are meant to help us choose between the bodily and the psychological criteria of personal identity do not help, because our concepts are not so sharp as to determine an answer in cases that do not arise in practice.

Johnston's provocative position is that although "we are not made to think of ourselves as ontological trash. Our basic pattern of self-regarding deliberation is seriously undermined if we are ontological trash. However, it is very difficult, at least within a materialist ontology, to see how we could be anything other than ontological trash." As to whether we are more than wholly material in nature, for Johnston this is a question that cannot be solved by entirely *a priori* methods "we philosophers are now under a clear obligation to learn a lot more science than the analysts of old deemed relevant. We need to get out of the armchair and look into things."

Hovering between the philosophy of mind and moral philosophy is the issue of free will. Russell and the positivists followed Hume in denying libertarian freedom and subscribing to compatibilism. G. E. Moore set up a particular line of compatibilist strategy known as the conditional analysis: *x could* have done otherwise means that he *would* have done otherwise if he had wanted to. A major issue in modern discussions concerns the adequacy of this account of freedom, in particular whether it can stand against the *consequence argument* that claims that you cannot choose to do otherwise if what you do is a strict consequence of events that precede your existence, as determinism claims. Ferenc Huoranszki investigates the state of these arguments.

Epistemology has been basic to philosophy at least since Plato, but its fundamental status for the empiricists and for science ensured that it would be vital to the analytic tradition. This centrality takes at least two forms. One is mainly concerned with conceptual analysis, and is the attempt to define or analyze the concept of knowledge. The other is more substantive and is concerned about what, if anything, we can know and how we can know it. Bryan Frances and Allan Hazlett examine analytic attempts to investigate both these issues, and also the question of the *value* of knowledge—why do we think it so important? They show that, on none of these questions have we yet managed to reach closure, despite a multitude of very sophisticated attempts.

The philosophy of perception is a perennial problem, especially in the empiricist tradition that gave birth to analytic philosophy. It figured centrally in the early twentieth century, mainly in the form of the sense-datum theory, but was largely driven off the stage by the ordinary language attack on that theory. But it has now returned with a vengeance, with even the sense-datum theory daring to show its face! Paul Snowdon takes the sense-datum theory as his starting position and investigates the various ways in which one might try to avoid it and defend a form of direct realism. He takes us carefully through this minefield, in the hope and expectation that this time round we can become genuinely clear on the issues and options.

We described in Part I the ethical tradition from Moore to Hare, and noted the so-called Great Expansion in analytic ethics that took place as part of the reaction against emotivism and Hare's meta-ethics. The reason for this is

probably best understood in the following way. Discussions falling under the label “ethics” or “moral philosophy” not unnaturally concerned *moral* right and wrong, but in fact the empiricist tradition from Hume, within which modern noncognitivists were working, was concerned with a much broader issue, namely whether anything could constitute a rational ground for action of any sort, including prudential and self-interested action. What is labeled the *fact-value* divide, or the *is-ought* problem is in fact the supposed great divide between *matters of fact* and *anything at all counting as a reason for action*. The lay moral skeptic, who might think of himself as an emotivist because he believes that *moral* values merely express subjective preferences, will probably think that acting on self-interest is entirely rationally grounded: the loss of rational foundations comes as one moves from selfishness to altruism. He has every reason, he will think, to avoid pain and seek happiness for himself, but no solid reason to take steps to do the same for a stranger on the other side of the world. But the radical skeptical view is that no fact—not even what pain feels like, or what wants and desires one possesses—constitute in themselves even *prima facie* rational grounds for action. This is the way that Hume’s chapter “Of the Influencing Motives of the Will” has been generally interpreted and it is certainly the message of Hare’s “Pain and Evil” and “Wanting: Some Pitfalls.” Desires or painful feelings may cause actions, but they constitute even *prima facie* grounds for action only if we choose or endorse them.

It seems to have been this deadlock over practical rationality in general that led philosophers into the whole realm of practical rationality and normativity, rather than restricting themselves to ethics in the traditional more narrow sense. It is the question of what sorts of things normative or practical reasons are that Ruth Chang takes up. She argues that the deadlock to be found in this debate between those who think that these reasons are desire based, those who think they are value based, and those who think that they are a hybrid of these two stems from a failure to distinguish between three questions. These concern the *content* of normative practical reasons, the *nature* of their normative force, and the *source* of that force. If these questions are clearly distinguished, she believes, we can make progress in understanding normativity.

Three main ethical theories dominate contemporary debates among analytical philosophers. According to consequentialists, the rightness or wrongness of a particular course of action depends not on the action itself, but on the consequences of that action. In contrast, for deontologists (in the Kantian tradition), it is actions themselves – or the intentions which lie behind them – which matter; some actions are intrinsically wrong. For virtue theorists (in the Aristotelian tradition), “the good” is essentially bound up with *good (human) lives*, and virtues are behavioural and emotional

tendencies which tend to promote human flourishing in those who possess them. Accordingly, the right course of action in a given situation is (roughly) what a virtuous person would do in that situation.

One familiar criticism of consequentialism is that it is overly lax. If no actions are intrinsically good or bad, and only consequences matter, there may be circumstances in which (say) killing or torturing an innocent person may be the right thing to do because the consequences acting thus are sufficiently good. In recent decades a different criticism of consequentialism has gained prominence: consequentialism is not only overly lax, it is also overly *demanding*. If you were to give just a small amount of money (\$1 say) to the right charity you will alleviate a great deal of suffering, e.g. you might save someone's sight, or prevent them dying of some awful disease. Since the consequences of giving just \$1 are so significant, and if consequences alone determine what is right, then you should give *more* than \$1, much more: you should keep on giving until you are very considerably poorer than you are now. In his chapter, Attila Tanyi outlines and assesses the recent discussions of this issue. Can a moral theory be rejected for demanding a great deal of us? Is it really true that consequentialism makes demands that other moral theories do not? If it does, what is the consequentialists' best response to the problem? Is a *multi-dimensional* consequentialism the answer? Can social institutions lower the demands morality makes on individuals? Or can *non-moral* reasons constrain what morality demands of us? The relationship between reason and normativity that Chang focuses on in her contribution turns out to be relevant here too.

It is probably fair to say that, until the arrival of Rawls on the scene, political philosophy was not a major area of analytic activity. Until the 1970s, the syllabus was dominated by historical figures such as Plato, Aristotle, Hobbes, Locke, and Mill. But the emergence of Rawls—a left-leaning liberal—and Nozick—a right-wing libertarian (sometimes dubbed “the rich man's Rawls”) brought political philosophy to the center of the contemporary stage. Andres Moles considers the question of the location and scope of the theory of justice, if one follows Rawls' principles. Does justice apply within given states, or to the world as one whole (“cosmopolitanism”)? Does it only concern the moral form of state institutions and policies, or should it inform the choices of individuals? One might rephrase the latter as the question, is Rawls' theory of justice simply political philosophy, or is it moral philosophy as well? Moles investigates G. A. Cohen's claim that it must be pushed into the individual as well as the political realm, and also defends a form of cosmopolitanism.

16

Mathematics and Logic

Mary Leng

Most introductions to the philosophy of mathematics have very little to say about the subject prior to the development of the various foundationalist projects in the late nineteenth and early twentieth century. There is a good reason for this. Some of the most pressing questions about the nature of mathematics cannot be properly considered in the absence of developments in logic (beyond syllogistic logic) that did not happen until the middle of the nineteenth century, with the development (by Boole, de Morgan, Pierce, and especially Frege), of a formal symbolism for logic and the introduction of a logic of relations (see Kneale and Kneale 1962). That the flourishing of the philosophy of mathematics as a serious discipline went hand in hand with the rise of analytic philosophy generally in the early twentieth century is no coincidence; both required developments in logic in order to make proper sense of their central questions.

Consider, for example, David Hume's bold empiricist statement that "All the objects of human reason or enquiry may naturally be divided into two kinds, to wit, *Relations of Ideas*, and *Matters of Fact*" (*Enquiry* IV, part i), and his placement of "the sciences of Geometry, Algebra, and Arithmetic" firmly into the first camp, as including propositions that are "discoverable by the mere operation of thought." What, for Hume, is the distinguishing feature of a proposition "discoverable by the mere operation of thought"? A look at the alternative category of matters of fact, may help here. For Hume, "The contrary of every matter of fact is still possible; because it can never imply a contradiction." Are we to take it, then, that Hume's paradigm examples of relations of ideas, "*That the square of the hypothenuse is equal to the square of the two sides*," and "*That three times five is equal to the half of thirty*" are such that their contraries imply contradictions? And if so, how is this to be established?

If we are right in reading Hume's discussion as implying that the truths of geometry, algebra, and arithmetic are knowable as "relations of ideas," by virtue of their negations being contradictory,¹ then Immanuel Kant's philosophy presents a challenge to this view. Kant certainly accepted this reading of Hume, arguing that Hume "erred severely" on precisely this point, and that recognizing the proper nature of mathematical truths would have

prevented Hume's rash dismissal of metaphysics. According to Kant in the 1783 *Prolegomena* [4: 272],

although [Hume] had by no means made a classification of propositions as formally and generally, or with the nomenclature, as I have here, it was nonetheless just as if he had said: Pure mathematics contains only *analytic* propositions, but metaphysics contains synthetic propositions *a priori*.

Hume of course wished to commit the supposed synthetic *a priori* propositions of metaphysics to the flames, for they "can contain nothing but sophistry and illusion" (Enquiry XII, part iii), but if Kant's diagnosis of the nature of mathematical truths is right, then the volumes of "divinity or school metaphysics" cannot be dismissed so easily.²

While Kant agreed with Hume that the truths of both are, as Hume suggested, discoverable by the mere operation of thought, this is not, he thought, because their negations imply contradictions. In Kant's view, though we can discover mathematical conclusions through discovering that their negations are contradictory, this is only against a backdrop of mathematical principles that are not themselves established in this way (*Critique* B15-16).

But if we look more closely we find that the concept of the sum of 7 and 5 contains nothing save the union of the two numbers into one, and in this no thought is being taken as to what that single number may be which combines both. The concept of 12 is by no means already thought in merely thinking this union of 7 and 5; and I may analyse my concept of such a possible sum as long as I please, still I shall never find the 12 in it. . .

Just as little is any fundamental proposition of pure geometry analytic. That the straight line between two points is the shortest, is a synthetic proposition. For my concept of *straight* contains nothing of *quantity*, but only of *quality*. (*Critique* B15-16)

For Kant, the propositions of arithmetic and geometry are knowable *a priori* not because they are analytic (and hence knowable by reflection on the meanings of the words involved, together with the principle of contradiction), but rather because their truth is graspable in intuition, by reflecting on the form of our experience of space (geometry) and time (arithmetic). In constructing geometric shapes or sequences of points in time in intuition, we can grasp the truth of principles of geometry and arithmetic independently of any particular experience.

In Kant's view, then, mathematics presents a challenge to empiricism and its neat division between relations of ideas (as "analytic" in Kant's terms, and knowable *a priori*) and matters of fact (as "synthetic" and knowable only *a*

posteriori). But how can we adjudicate between Hume and Kant? A major difficulty is in characterizing the supposed category of “relations of ideas” or “analytic” truths. For Kant, analytic truths are strictly those in which the predicate belongs to the subject (*Critique* A6/B10). However, Kant also uses the notion, which appears to be implicit in Hume’s discussion, of following from the principle of contradiction (in discussing our knowledge that a body is extended): “I have already in the concept of body all the conditions required for my judgment. I have only to extract from it, in accordance with the principle of contradiction, the required predicate” (ibid., B12). These two criteria are, however, distinct, and at any rate cannot be applied to those many judgments that are not of subject-predicate form (what, for example, of claims involving polyadic relations rather than monadic predicates?).

The rise of modern logic meant that progress could be made in navigating the debate between Kant and Hume, as it allowed for a sharpening of the notion of an analytic truth that allowed the notion of truth in virtue of meaning to be extended to statements that are not in subject-predicate form. Thus, having invented the predicate calculus with his *Begriffsschrift* (1879), in order to consider the question of “how far one could get in arithmetic by inferences alone, supported only by the laws of thought that transcend all particulars” (*Begr* IV, p. 48), Frege was able to sketch in his *Grundlagen der Arithmetik* (1884) a foundation for arithmetic in logic, which he hoped “made it probable that the laws of arithmetic are analytic judgements” (*Grundlagen*, p. 102), and which would secure the certainty of arithmetic by ruling out “the possibility that we may still in the end encounter a contradiction which brings the whole edifice down in ruins” (ibid., p. ix). Frege filled out the details of this sketch of his logicism in the two volumes of his *Grundgesetze der Arithmetik* (Basic Laws of Arithmetic; 1893/1903), which provide a derivation of the Dedekind-Peano axioms for the natural numbers from a number of basic laws of (second-order) logic. Thus his development of predicate logic enabled Frege to provide a sense in which it could be argued that arithmetic (though not geometry—on this Frege agreed with Kant) was analytic.

Alas, Frege’s logicist account of arithmetic as analytic was fatally flawed, as he discovered when Bertrand Russell wrote to him in 1902, on the eve of the publication of the second volume of the *Grundgesetze*. “I find myself in complete agreement with you in all essentials,” wrote Russell, adding that there was “just one point where I have encountered a difficulty”:

You state (p. 17) that a function, too, can act as the indeterminate element. This I formerly believed, but now this view seems doubtful to me because of the following contradiction: Let w be the predicate: to be a predicate that cannot be predicated of itself. Can w be predicated of itself? From each answer its opposite follows. Therefore we must conclude that w is

not a predicate. Likewise there is no class (as a totality) of those classes which, each taken as a totality, do not belong to themselves. From this I conclude that under certain circumstances a definable collection does not form a totality. (Russell 1902, pp. 124–5)

Despite the high esteem in which Russell held Frege's work, this discovery of what would come to be known as "Russell's Paradox" was devastating to Frege's project. Frege had assumed in one of his basic laws (Basic Law V) that every predicate has a corresponding extension (a collection containing precisely those objects that the predicate can be truly predicated of). But Russell's predicate w , a predicate of predicates, can truly be applied to (i.e. predicated of) precisely those predicates that cannot be predicated of themselves. It follows from this that w can be predicated of itself if and only if it cannot, a contradiction. Frege had hoped to found arithmetic in basic laws of logic whose truth would be so transparent once one had grasped the meanings of their logical terms that they themselves needed no further foundation, and thus establish mathematical knowledge as grounded a priori in laws of thought. But Basic Law V turned out not simply to be false, but to be contradictory, destroying the foundation Frege had worked so hard to provide, and bringing his own edifice down in ruins.

Frege realized the significance of Russell's "just one point" immediately, responding that

Your discovery of the contradiction caused me the greatest surprise and, I would almost say, consternation, since it has shaken the basis on which I intended to build arithmetic. It seems, then, that transforming the generalization of an equality into an equality of courses-of-values (§9 of my *Grundgesetze*) is not always permitted, that my Rule V (§20, p. 36) is false, and that my explanations in §31 are not sufficient to ensure that my combinations of signs have a meaning in all cases. I must reflect further on the matter. It is all the more serious since, with the loss of my Rule V, not only the foundation of my arithmetic, but also the sole possible foundation of arithmetic, seems to vanish. Yet, I should think, it must be possible to set up conditions for the transformation of the generalization of an equality into an equality of courses-of-values such that the essentials of my proofs remain intact. In any case your discovery is very remarkable and will perhaps result in a great advance in logic, unwelcome as it may seem at first glance. . . .

The second volume of my *Grundgesetze* is to appear shortly. I shall no doubt have to add an appendix in which your discovery is taken into account. If only I already had the right point of view for that! (Frege 1902, pp. 127–8)

Frege never did find a satisfactory resolution to the paradox within his system, and ultimately abandoned his logicism. Russell, on the other hand, continued to defend the reduction of mathematics to logic, and presented an alternative foundation for mathematics in *Principia Mathematica* (coauthored with Alfred North Whitehead). But his foundation, based as it was on the ramified theory of types that he had introduced to replace Frege's faulty assumptions about extensions, required the assumption of axioms including the axioms of infinity and replacement that were far from logical truths, and thus fell short of the logicist aim of establishing the analyticity of arithmetic via a reduction of arithmetic truths to logical truths plus definitions.

With the collapse of Frege's logicist project one might expect a reversion to a Kantian view of mathematics as synthetic a priori. While Frege's development of the predicate calculus provided the conceptual resources required for a workable notion of "analyticity" (i.e. derivable from logical truths together with definitions), he had not succeeded in displaying arithmetic as analytic in this sense. Alternative reductions such as that of Russell's *Principia*, or reductions of numbers to sets as characterized by the Zermelo-Fraenkel axioms for set theory, were able to found our concept of number in a concept of set, but given the nonlogical character of the assumptions such theories had to make about sets, it appeared that all that had been achieved was a foundation of one branch of mathematics in another, not a reduction to logic in any epistemically important sense. But despite the apparent failure of logicism, Kantianism could not succeed "by default." Important developments in mathematics, and in particular the development of infinitary set theory, presented a separate challenge to the Kantian view of arithmetic as grounded in our intuition of time.

Over the course of the nineteenth century, starting with the work of Cauchy and Bolzano, and developed by Weierstrass, Cantor, and Dedekind, a movement of rigorization had developed in mathematical analysis. Analysis is the mathematical theory of continuous change, and as such includes the calculus as developed independently by Newton and Leibniz in the seventeenth century. While the development of the calculus was a magnificent achievement, the early theory involved conceptual difficulties that had not gone unnoticed. Indeed, in 1734 the philosopher George Berkeley published his pamphlet *The Analyst*³ where he ridiculed the use of infinitesimal quantities in the calculus, asking

And what are these Fluxions? The Velocities of evanescent Increments? And what are these same evanescent Increments? They are neither finite Quantities nor Quantities infinitely small, nor yet nothing. May we not call them the Ghosts of departed Quantities? (p. 18)

Berkeley had a point: the calculus as developed by Newton and Leibniz, though practically extremely fruitful, had at its roots a notion of infinitesimal that was conceptually a mess. The developments of the nineteenth century, and in particular Karl Weierstrass's epsilon-delta notion of continuity, rectified these problems by dispensing with infinitesimal quantities in analysis, replacing talk of infinitesimals with talk of what happens to infinite sequences of finite quantities as their terms get ever closer to zero. As well as dispensing with infinitesimals, the newly rigorized calculus dispensed with geometrical intuition in favor of explicit definitions in terms of arithmetic concepts.⁴

The rigorization of the calculus removed the problematic infinitesimals, but in doing so made ever clearer the dependence of the central concepts of analysis on infinitely large collections. This is particularly clear in the various proposed definitions of real numbers. In his *Cours d'Analyse* (1821) Augustin-Louis Cauchy noted that irrational (by which he meant, "real") numbers could be identified as "the limit of diverse fractions which furnish closer and closer approximate values of it," an idea that was later (1872) adopted by Cantor to provide a definition of real numbers (as limits of Cauchy sequences of rationals). An alternative definition of real number was also published in 1872, by Richard Dedekind. Dedekind's basic idea was to think of real numbers as divisions in the real line. He introduced the notion of a "cut" in the rationals—a division of the rational numbers into two disjoint and nonempty sets (A_1 , A_2), where every number in the first of these sets is less than any number in the second. If A_1 has a largest member, or A_2 has a smallest, then the cut determines a rational number (that being the largest member of A_1 or smallest member of A_2). But if A_1 has no largest member and A_2 no smallest, then the cut is defined to be an irrational number (intuitively thought of as the unique point standing between all the members of A_1 and all the members of A_2). Whether we adopt Cantor's, or Dedekind's, account of real number, in either case real numbers are defined in terms of infinite sets. As such, in Morris Kline's view, the "irrational number, logically defined, is an intellectual monster" (1972, p. 987).

Recognizing the need for a theory of infinite sets to underpin the new developments in analysis, Georg Cantor developed transfinite set theory in a series of papers from 1874 onwards. Cantor's work opened up a surprising new realm of mathematical enquiry. Cantor introduced the notion of the cardinality of a set, defining two sets as having the same cardinality if and only if they can be put in one-one correspondence with each other. From this he was able to prove that the rational numbers and algebraic numbers had the same cardinality as the natural numbers, but that the real numbers had a greater cardinality than all of these, and, in general, that the cardinality of the power set of a set A is greater than the cardinality of A . What Cantor had shown,

then, was that rather than a single category of “the infinite,” there was an unlimited stock of ever larger infinite sizes of sets beyond the smallest “countable” infinity of natural numbers (which Cantor labelled \aleph_0).

These developments were problematic for a Kantian view of arithmetic. So long as arithmetic involved only the finite counting numbers, it was plausible to think of our grasp of the truths of arithmetic as grounded in our temporal intuition. For any finite cardinal number n we can conceive (if we ignore limitations of time, materials, and attention) of our being able to construct that number in intuition, in the sense, for example, of writing down a series of successive strokes $|||| \dots |$, perhaps one every second and stopping after n seconds. But even the smallest infinite cardinality, \aleph_0 , cannot be constructed in time in this way, and even if we could convince ourselves that we could “in principle” construct a set of size \aleph_0 (perhaps by writing down a first stroke in one second, a second in half a second, a third in a quarter of a second, and so on, so that by two seconds an infinite number of tasks will have been completed each taking half the time of the previous task), we would certainly not be able to construct the larger infinite sets in this way. So if our knowledge of arithmetic is grounded in temporal intuition, then it is entirely unclear how Cantor’s transfinite arithmetic can be counted as knowledge.

One response to this situation, from within the Kantian camp, is to reject the new infinitary mathematics. This was the line taken by intuitionists such as L. E. J. Brouwer. Brouwer, as a Kantian, held that meaningful mathematics was grounded in our constructions in intuition. He thus abandoned transfinite mathematics. But that was not all that had to go. Brouwer saw that many results of classical mathematics were ungrounded once one viewed the truths of mathematics as constituted by our human mental constructions. In particular, the law of the excluded middle ($p \vee \neg p$) is not valid if one holds that a proposition’s being true consists in our having constructed its proof. There are many mathematical propositions that have neither been proved nor refuted (take, for example, Goldbach’s conjecture that every even number greater than two is a sum of two prime numbers). Since we have neither constructed a proof of Goldbach’s conjecture, nor constructed a proof of its absurdity, neither p nor $\neg p$ is true, on Brouwer’s view of mathematical truth. Abandoning this law would have been even more drastic than the abandonment of Cantor’s new set theory, as it would have meant that long established mathematical proofs needed to be revisited.

Brouwer’s intuitionism was not simply a negative program. Brouwer realized the felt need for infinitary mathematics in grounding the newly rigorized analysis, and made significant progress in developing a constructive alternative to the classical theory of the continuum. However, the majority of mathematicians still preferred classical forms of reasoning, and classical analysis. Constructive mathematics, including the intuitionistic theory of the

continuum, presented an interesting new branch of mathematics, but for the most part mathematicians, unphased by Brouwer's Kantian scruples, continued to reason classically, and to make use of standard analysis, as before.

An alternative approach within the Kantian vein tries to preserve classical mathematics despite its lack of grounding in intuition, thus justifying the continued use of classical reasoning and infinitary mathematics. In his 1925 lecture, "On the Infinite," David Hilbert noted the tension between the use of infinitary mathematics and the lack of an intuitive ground for our concept of infinity. On the one hand, "the infinite is nowhere to be found in reality. It neither exists in nature nor provides a legitimate basis for rational thought" (p. 201), yet "mathematical analysis is a symphony of the infinite" (*ibid.*, p. 187). The lack of an intuitive ground was not a mere philosophical concern arising out of an adherence to Kant's philosophy. As Hilbert noted, Cantor's new set theory in its early developments was littered with contradictions, including Cantor's paradox (of the set of all sets: assuming there is such a set, it should be maximal, yet its power set would, like any power set, have to have a strictly greater cardinality, and so contain even more sets), and most crucially, Russell's paradox. Yet, rather than following Brouwer in abandoning the infinite, Hilbert still wished to preserve infinitary mathematics, outlining the following goals:

1. Wherever there is any hope of salvage, we will carefully investigate fruitful definitions and deductive methods. We will nurse them, strengthen them, and make them useful. No one shall drive us out of the paradise which Cantor has created for us.
2. We must establish throughout mathematics the same certitude for our deductions as exists in ordinary elementary number theory, which no one doubts and where contradictions and paradoxes arise only through our own carelessness. (Hilbert 1926, p. 191)

How did Hilbert hope to preserve the paradise of infinitary mathematics in the light of his own belief in the meaninglessness of its central concepts? Following Kant, Hilbert held that the objects of mathematics are not part of logic, but "certain extralogical concrete objects which are intuited as directly experienced prior to all thinking" (*ibid.*, p. 192). Rather than mental constructions, though, Hilbert thought of these objects as the symbols we can write down, starting specifically with

"the numerical symbols 1, 11, 111, 1111" (*ibid.*)

which can then be abbreviated, and symbols such as "+," "=", ">," to communicate the ideas of the concatenation of two strings of symbols, of two strings

being the same length, and of the first string being longer than the second. The meaningful statements of number theory are those claims about strokes whose truth can be established in finite time. That is, equations and inequations; bounded existential quantifications (asserting the existence of a number with a given finitely checkable property where that number is smaller than some finite bound); and finitary general claims where there is a general procedure for establishing each instance (e.g. all instances of " $a + 1 = 1 + a$ " where " a " is replaced by a numerical symbol). Anything in the theory of numbers falling outside of this category (effectively, anything beyond primitive recursive arithmetic) would not count as meaningful.

This ruled out Cantor's arithmetic of cardinal numbers, but as with Brouwer, Hilbert's restrictions to what could be grounded in finite manipulations of finitary symbols led him to question some more traditional mathematical reasoning too. If we view " $a + 1 = 1 + a$ " as is usual, as a universally quantified claim about the natural numbers, then its negation is the existentially quantified (and false) claim, "there is a natural number a such that $a + 1 \neq 1 + a$." But on Hilbert's reading of " $a + 1 = 1 + a$," nothing is asserted until a particular numeral is replaced for " a ": "this statement cannot be interpreted as a conjunction of infinitely many numerical equations by means of 'and' but only as a hypothetical judgment which asserts something for the case when a numerical symbol is given" (ibid., p. 194) As a result, " $a + 1 = 1 + a$ " is "*incapable of negation*," and so Hilbert rejects the unrestricted use of the law of excluded middle in the finitary core of mathematics. Some finitarily meaningful statements do not have finitarily meaningful negations, so " $p \vee \neg p$ " is not universally true for finitary p .

Unsurprisingly then, given their shared Kantian roots, Hilbert and Brouwer were led to similar restrictions on meaningful mathematics, with both of them rejecting the meaningfulness of Cantor's transfinite set theory, and both of them questioning the law of the excluded middle. But as we have said, Hilbert wished to preserve classical, and transfinite, mathematics, not to abandon it. Hilbert's solution was an instrumentalist one: he adopted infinitary mathematics as a strictly meaningless formal tool, in order to aide our reasoning about the finitary, meaningful core of mathematics. Setting aside finitary scruples meant Hilbert could treat statements such as " $a + 1 = 1 + a$ " as if they were genuine universal generalizations, with meaningful negations, and thus preserve the use of the law of the excluded middle in mathematics. The full force of Cantor's transfinite set theory could also be preserved (insofar as the paradoxical elements had been banished and the techniques of reasoning been clarified and formalized), not as a meaningful and true theory about the infinite, but as a useful formal tool whose value was ultimately in its role in enabling us to establish meaningful, finitary truths. Hilbert talked of these additional, strictly meaningless parts of mathematics

as “ideal statements,” in analogy with the use of ideal elements such as points at infinity or imaginary numbers to smooth out our theories of geometry and algebra respectively.

Hilbert’s formalist solution to the Kantian predicament is ingenious. Kantian scruples about meaning are acknowledged, yet the full force of classical and transfinite mathematics is retained, holding that not all parts of our theories have to be meaningful or true in order to be useful. “There is,” Hilbert noted,

just one condition, albeit an absolutely necessary one, connected with the method of ideal elements. That condition is a *proof of consistency*, for the extension of a domain by the addition of ideal elements is legitimate only if the extension does not cause contradictions to appear in the old, narrower domain, or, in other words, only if the relations that obtain among the old structures when the ideal structures are deleted are always valid in the old domain. (1926, p. 199)

In order to justify our reliance on strictly meaningless ideal mathematics, Hilbert saw that we needed to be confident that our reasoning using ideal elements, if carried out correctly, would not enable us to prove false claims about the finitary, meaningful core of mathematics. This amounted to proving the consistency of the theory that resulted from adding ideal elements to the finitary core of arithmetic. As Hilbert noted, the claim that a theory is consistent is a finitarily meaningful one (given an ordering on proofs (e.g. lexicographic), it is equivalent to the finitary general claim that, “the a -th proof does not end in $1 \neq 1$,” understood as being held to be true when a is replaced by any numerical symbol). Hilbert, having developed proof theory in order to complete his program of justifying our use of ideal mathematics, thought that a finitary proof of this finitary general claim was well within his grasps. In fact, in his lecture he asserts that “we can, as a matter of fact, prove that it is impossible to get a proof which has that formula as its last formula” (ibid., p. 200), though no proof is actually provided.

Unfortunately, Hilbert’s optimism about the prospects for finding a finitary consistency proof for infinitary mathematics was mistaken, as was shown by Kurt Gödel in his 1931 paper, “On formally undecidable propositions of *Principia Mathematica* and related systems I.” Gödel’s paper contained two “incompleteness” theorems of deep significance for the philosophy of mathematics. In his first incompleteness theorem, Gödel showed that, in any axiomatic theory T powerful enough to contain primitive recursive arithmetic (i.e. Hilbert’s finitary arithmetic), if T is consistent then there is a sentence G in the language of T such that neither G nor $\neg G$ is derivable in T . This in itself was entirely unexpected—indeed, in his 1925 lecture Hilbert asserted of “the

thesis that every mathematical problem is solvable” that “we are all convinced that it really is so.”

In fact, one of the principal attractions of tackling a mathematical problem is that we always hear this cry within us: There is the problem, find the answer; you can find it just by thinking, for there is no *ignorabimus* in mathematics. (Hilbert 1926, p. 200)

Gödel’s first theorem showed that this was an error. Whatever axioms we adopt for a mathematical theory, if those axioms are consistent and if they are sufficient to include basic arithmetic, then there will be some mathematical questions whose answers we cannot find by derivation from those axioms.

But the real disaster for Hilbert’s program of justifying our use of infinitary mathematics was a result of Gödel’s second incompleteness theorem, itself a corollary of the first. The second theorem states that, for a mathematical theory *T* as above, if *T* is consistent then it is unable to prove its own consistency. Hilbert hoped for a proof of the consistency of infinitary mathematics using its meaningful, finitary core. But Hilbert’s infinitary mathematics included finitary mathematics as a component. So a finitary proof of the consistency of infinitary mathematics would also be a finitary proof of the consistency of finitary mathematics, which is precisely what the second incompleteness theorem proves to be impossible. As Russell’s paradox did for Frege’s logicism, once again a result in logic had brought an attempt to provide a philosophical foundation for mathematics to its knees. Hilbert’s program never recovered from this blow.

As of 1931, then, the position in the philosophy of mathematics looked pretty desperate. Despite the immense promise that the developments in logic had had for providing a firm foundation for mathematics, those same developments had worked to scupper the leading foundationalist programs. Frege’s logicist attempt to show arithmetic to be analytic had been destroyed by the discovery of Russell’s paradox, and Russell’s own reduction of mathematics to set theory was further challenged by Gödel’s first incompleteness theorem, which showed that some mathematical truths would not be derivable from Russell’s—or anyone’s—axioms, so long as those axioms were consistent. On the other hand, Hilbert’s Kantian attempt to preserve the new infinitary mathematics despite its lack of a grounding in constructions in intuition had been abandoned in the light of Gödel’s second incompleteness theorem. Brouwer’s intuitionism survived these attacks, but remained for most an unattractive approach, owing to the radical revisions it required in the practice of mathematics. The whole issue of the nature of mathematics needed a radical rethink.

Gödel's first incompleteness theorem, in particular, required a shake up in the way people thought about the truths of mathematics. Prior to this theorem, a natural thought was that truth in mathematics was quite different from truth elsewhere. Whereas truths about the physical world involved agreement with an external reality, it was natural to think of truth in mathematics as an internal matter, something that could be found, in Hilbert's words, "just by thinking," once one had grasped the concepts of mathematics (and perhaps, if Kant is right, carried out some intuitive constructions). One way of cashing out this idea was to think of mathematical truths as being true in virtue of following from our mathematical assumptions. But if by "following from" we mean "being derivable from," then Gödel's first theorem presents a staunch challenge to this picture: whatever axioms for arithmetic we choose, so long as they are consistent there will be more truths about the natural numbers than are derivable from those axioms.

A natural response to this predicament is (arguably) Gödel's own: to reject the identification of mathematical truth with following from axioms in favor of a more substantial notion that makes truth-in-mathematics more closely analogous to truth *simpliciter*. On the face of it, the truths of mathematics appear to be about a realm of mathematical objects (the truths of Peano arithmetic being about the natural numbers; the truths of Zermelo–Fraenkel set theory with the axiom of choice (ZFC) being about the sets). If such objects have an independent existence, then perhaps we should not be surprised that our theories of these objects fall short in answering all questions about them, just as we are not surprised to find that there are empirical truths that are not answered by our empirical scientific theories? On a substantial (Platonist) understanding of mathematical objects existing independently of our theories about those objects, the truths of mathematics are true in virtue of what those objects are really like, not in virtue of whatever it is our theories have to say about those objects. Thus Gödel had no problem with holding that a theory's Gödel sentence was true *of the natural numbers* even though unprovable within that theory, since there is a well-determined reality of natural numbers that makes that sentence true. And similarly with set theory. Writing about Cantor's continuum hypothesis (which Gödel rightly suspected to be independent of the ZFC axioms), Gödel argued that

the set-theoretical concepts and theorems describe some well-determined reality, in which Cantor's conjecture must be either true or false. Hence its undecidability from the axioms being assumed today can only mean that these axioms do not contain a complete description of that reality. (1947, p. 476)

Taking seriously the Platonist picture of mathematical truths as true of a realm of independently existing mathematical objects suggests a picture of mathematics as a science in its own right, with its own subject matter. Gödel's discussion is certainly suggestive of such a view. Accordingly, Gödel looks for ways of finding out about mathematical objects that parallel our means for finding out about physical objects. Rather than a sharp distinction between the *a priori* methods of mathematics and the *a posteriori* methods of natural science, Gödel points to quasi-empirical methods by which we can form hypotheses about mathematical objects (justifying mathematical axioms by their fruitful consequences, for example). And crucially, for Gödel, in justifying our mathematical beliefs we do sometimes rely on experience, albeit not sensory experience. Thus, according to Gödel,

Despite their remoteness from sense experience, we do have something like a perception also of the objects of set theory, as is seen from the fact that the axioms force themselves upon us as being true. I don't see any reason why we should have less confidence in this kind of perception, i.e., in mathematical intuition, than in sense perception. . . . (ibid., p. 484)

Taking seriously the idea of a realm of mathematical objects thus pushes against the sharp distinction between mathematical knowledge and knowledge of physical objects.

This distinction was questioned even further by W. V. Quine (1951), as part of his famous attack on the "two dogmas" at the root of empiricism. The first of these dogmas is the analytic/synthetic distinction, which Quine argues has not been drawn in a way that could underpin its supposed role in empiricist epistemology, in particular as marking out some of our knowledge as nonempirical, knowable in virtue of the meanings of words. The second is reductionism, the idea that a sentence taken in isolation can be reduced to a collection of observable consequences, a dogma that, Quine says, "survives in the supposition that each statement, taken in isolation from its fellows, can admit of confirmation or infirmation at all" (1951, p. 41). Against this Quine presses his famous "countersuggestion," that "our statements about the external world face the tribunal of sense experience not individually but only as a corporate body" (ibid.), with radical consequences for our view of the status of mathematical knowledge.

If Quine is right in his rejection of the analytic/synthetic distinction, then the project of separating off mathematical knowledge from "empirical" knowledge by virtue of its analyticity is mistaken. Instead, in Quine's view, our knowledge of the truths of mathematics is as rooted in empirical experience as the rest of our knowledge. Our empirical scientific theories are formulated in mathematical terms: we cannot even state their laws without reference to

functions and real numbers. But if “our statements about the external world face the tribunal of sense experience not individually but only as a corporate body” then the mathematical portions of our scientific theories are tested against experience along with those scientific theories, and are confirmed to the extent that those theories as a whole are confirmed by their empirical successes. Far from having a special status as a priori due to the analytic character of mathematical truths, mathematical knowledge is as grounded in experience as the rest of our knowledge, receiving confirmation as part of its place in our entire web of belief. Mathematical knowledge is not, as Gödel suggested, *analogous* to empirical knowledge; rather, mathematical knowledge *is* empirical knowledge, with mathematical truths being confirmed in just the same way as truths about the empirical world are confirmed.

Can this be right, though? Do mathematical objects not have a special status that suggests that they should be insulated from confirmation in this manner? Mathematical truths are meant to be atemporal, unchanging, mind-independent. They are meant to be necessary, not dependent on the contingencies of the world we happen to find ourselves in. To underpin truths of this sort, mathematical objects are generally thought of as *abstract*, where “abstract” is characterized negatively: nonspatiotemporal; acausal; mind-independent. And this raises the question, how could the empirical evidence we receive through the senses speak to the existence and nature of objects of *that* sort? This is the key question of Paul Benacerraf’s influential (1973) paper, “Mathematical Truth.” If we take mathematical truth seriously as a genuine species of truth (not, for example, reducible to “following from consistent axioms”), then we appear to be compelled to accept, as Gödel does, a realm of mathematical objects about which our mathematical theories assert truths. But our best characterization of these objects (as nonspatiotemporal, acausal, and mind-independent) makes it a mystery as to how we, as spatiotemporally located beings whose information about the external world is received through the senses, could gain knowledge of truths about such objects.

Benacerraf had already, in his (1965) paper, “What numbers could not be,” presented a significant difficulty for the Platonist view of mathematical objects. If mathematical truths are truths about a realm of objects, there must be some fact of the matter about precisely which objects these objects are. Indeed, Frege viewed it as a requirement on any definition of number that it should provide the means to answer any grammatically correct identity question about numbers (such as, famously, “Is the number 2 identical with Julius Caesar?”). It was this requirement that led Frege to identify numbers with extensions of concepts, bringing in his ill-fated Basic Law V to explain when the extensions of two concepts are identical. Frege’s account was of course inconsistent, but since Frege, alternative accounts of numbers as sets had been provided, for example, by Russell from the perspective of the ramified theory

of types, and by Zermelo and von Neumann from the perspective of ZFC, and as far as we know, these accounts, unlike Frege's, are consistent. The problem Benacerraf notes is that we appear now to have an embarrassment of riches. Numerous reductions of numbers to sets, or to other objects, are available, and it appears that nothing in our concept of number enables us to choose between them. Against the Gödelian Platonist view, then, Benacerraf suggests that, "despite appearances, arithmetic is not a science concerned with particular objects—the numbers. The search for which independently identifiable particular objects the numbers really are (sets? Julius Caesars?) is a misguided one" (1965, p. 291). By itself, Benacerraf's (1965) paper presents a puzzle for the standard Platonist view of mathematical truths as truths about a realm of objects. Combine this with the concerns of Benacerraf (1973), and Gödelian Platonism appears to be in major difficulties.

Quine's holistic epistemology and his placement of mathematical truths alongside empirical truths, and Benacerraf's emphasis on the rather special, and strange, nature of mathematical objects, thus pull in two different directions. On the one hand, by rejecting the analytic/synthetic distinction and taking a holistic view of confirmation, Quine is able to offer an epistemology for Platonism that places mathematical objects along a continuum that ranges from brick houses to electrons to classes (sets), all of which receive confirmation through their role in a smooth theory of our sensory experiences. On the other hand, Benacerraf points to important differences that appear to speak against viewing mathematical objects as on a par with theoretical posits such as electrons. The view of mathematical theories as about specific objects seems problematic in the light of the existence of multiple acceptable systems of objects that can "do the job" of (for example) the natural numbers. And the view of mathematical truths as receiving empirical confirmation is hard to square with an understanding of their objects as independently existing *abstracta*: how could empirical evidence speak in favor of, or against, claims about such things, given their causal isolation from the physical world? How could truths about such objects leave a trace in the world we inhabit? It is hard to see how empirical evidence could give us any confidence in the existence of such things, despite Quine's holistic countersuggestion.

This tension, between Quine's holistic placement of mathematical objects alongside the theoretical objects of empirical science and Benacerraf's emphasis on the peculiar nature of mathematical objects and special status of mathematical truth, runs through contemporary debates in the philosophy of mathematics. I will consider neo-logicism, fictionalism, structuralism (modal and ante rem), and full-blooded Platonism in relation to these debates.

Just as Frege stood firm against Kant as concerns the status of the truths of arithmetic as analytic rather than synthetic, the contemporary neo-logicians (such as Crispin Wright and Bob Hale) stand firm against Quine's attack on the

analytic/synthetic distinction, holding that there remains a workable distinction and that we can uphold the view of arithmetic as analytic (see the papers in Hale and Wright 2001). By pressing the analytic nature of mathematical truths, neo-logicists suggest a traditional answer to Benacerraf's knowledge problem: mathematical truths are known through reflection on language and the meanings of our words.⁵ Neo-logicists avoid both Russell's paradox and the difficulties of reducing mathematics to a set theory that makes use of axioms that are not truths of logic by rejecting Frege's requirement to give an account that states explicitly which objects the numbers are. Frege himself showed (in a result that has come to be known as "Frege's Theorem") how the Peano axioms for arithmetic could be derived from "Hume's Principle," the claim that the number of Fs is identical with the number of Gs if and only if the Fs and the Gs are equinumerous, but Frege rejected Hume's principle as a definition number because it failed to solve the Julius Caesar problem. Neo-logicists, however, are happy to abandon the requirement to say precisely which objects the numbers are (perhaps taking a lesson from Benacerraf 1965), and to take Hume's Principle (which they take to be analytic of our concept of number) as their foundation from which the standard axioms of arithmetic are to be derived.

In carrying out their defense of the analyticity of mathematics, neo-logicists must face a number of challenges. If we accept Hume's Principle as a truth of second-order logic, then the existence of the natural numbers follows as a consequence of Hume's Principle. And neo-logicists suggest we have strong grounds to accept Hume's Principle, as characterizing our concept of number. But there are principles structurally similar to Hume's principle that have very different consequences for ontology. For example, George Boolos' "Parity Principle," according to which "the parity of F is equal to the parity of G if and only if F and G differ evenly"⁶ is true only of finite domains. The Parity Principle is thus true in no domains containing the natural numbers. Neo-logicists, then, owe an explanation of why we should adopt Hume's Principle (as analytic of our concept of number) rather than the Parity Principle (as analytic of our concept of parity), given that at most one of these principles can be adopted. Admittedly, our concept of parity is less entrenched than our concept of number, but is familiarity reason enough to adopt Hume's Principle rather than the Parity Principle as a conceptual truth?

Another challenge, of interest in what follows, is for neo-logicists to justify their use of second-order logic. The formal statement of Hume's Principle requires second-order quantification, in order to state what it is for concepts F and G to be equinumerous (i.e. "*there is a function ϕ such that ϕ is a one-one correspondence between the Fs and the Gs*"). The Peano Axioms are consequences, in second-order logic, of Hume's Principle. But there are question

marks over the acceptability of second-order logic *as logic*. W. V. Quine (1970) famously referred to second-order logic as “set theory in sheep’s clothing,” since quantifying into predicate position seems dangerously close to quantifying over sets. If all that Frege’s Theorem shows is that arithmetic can be reduced to Hume’s Principle *plus set theory*, then it is unclear what epistemic advantage neo-logicism has over, say, Russell’s reduction of arithmetic. To the extent that our theory of the sets will depend on axioms that are not themselves truths of logic, it appears that all we have is a reduction of one branch of mathematics to another. I will not consider here the case for taking second-order logic as logic, and for rejecting its identification with set theory. Rather, I would just like to note the importance of this issue for neo-logicism, not least because, as we will see, the same issue is faced by other contemporary philosophical accounts of mathematics.

Moving to philosophical accounts of mathematics that broadly accept Quine’s attack on the analytic/synthetic distinction and his resulting “naturalistic” outlook, which views the question of the existence of mathematical objects as a question for natural science to answer, accounts can be distinguished on the basis of whether or not they take the truth of (some) mathematics to be confirmed by its use in empirical science. Aside from Quine’s own discussions of the position of mathematics in our empirical theories, Hilary Putnam (1971, 1975) presses the indispensability of mathematics in formulating our scientific theories, stating the realist consequences of Quine’s philosophy most explicitly in what has come to be known as the Quine-Putman indispensability argument:

quantification over mathematical entities is indispensable for science, both formal and physical; therefore we should accept such quantification; but this commits us to accepting the existence of the mathematical entities in question. This type of argument stems, of course, from Quine, who has for years stressed both the indispensability of quantification over mathematical entities and the intellectual dishonesty of denying the existence of what one daily presupposes. (Putnam 1971, p. 347)

Philosophers influenced by naturalism agree with Quine and Putnam that we should for the most part accept our best scientific theories, holding with Quine that “there is no place for a prior philosophy” that can come to any better conclusions about the nature of reality than those we have been led to in empirical science. But precisely what is involved in accepting those theories remains in question. Is Putnam right that quantification over mathematical entities is indispensable for science and therefore should be accepted? And if so, does that mean accepting the existence of the mathematical entities quantified over?

Until relatively recently, the strongest challenge to the Quine-Putnam indispensability argument from within a broadly naturalistic perspective has been Hartry Field's fictionalism. Field (1980) agrees with Quine and Putnam that *if* quantification over mathematical entities is indispensable to science then we should accept that quantification and therefore accept the existence of mathematical objects. But Field argues that we can dispense with such quantification, rewriting the laws of our scientific theories so that they do not quantify over any mathematical objects (Field calls such theories "nominalistically stated," as opposed to the "platonistic" theories of ordinary science). Field's proposal for rewriting our scientific theories is not as radical as it sounds; he does not require us to abandon the use of mathematics in empirical science. Rather, we can justify our reliance on our ordinary (platonistic) scientific theories since these, he argues, are *conservative extensions* of the nominalistically stated scientific theories whose truth we take to be confirmed by the success of science. Informally, a platonistic scientific theory *S* is a conservative extension of a nominalistically stated theory *N* if any nominalistically stated consequence of *S* is already a consequence of the theory *N*. If *S* is a platonistic theory that we know to be a conservative extension of *N*, and we believe that *N* is (approximately) true, then if *O* is an observable consequence that we derive from *S*, we have reason to believe *O* even if we do not believe that *S* is true, since (by conservativeness) we know that *O* is also a consequence of *N*, and so true if *N* is true. In this case it is not intellectually dishonest to rely on *S* to derive predictions even if we do not believe the mathematical portions of *S*, since we have (via our account of the conservativeness of *S* over *N*) an explanation of why *O* is likely to be true even if *S* is not. As a result, it is appropriate to view our mathematical theories as merely useful fictions, rather than literally true parts of our best-confirmed scientific theories.

Justifying the various components of Field's challenge to the indispensability argument is hard work. Field sketches a nominalized version of Newtonian gravitational theory in *Science without Numbers*, but successful completion of his project would require extending Field's nominalization strategy to contemporary scientific theories, a task that many critics doubt is possible given the form of those theories. Newton's laws can be thought of as ultimately about spacetime points and their properties, so Field's nominalization of Newtonian science takes the basic objects to be spacetime points and expresses properties of these in nonmathematical terms. But our best current scientific theories are expressed as phase space theories, whose basic objects are the possible states of a physical system. Even if Field could extend the strategy of his nominalization of Newtonian science to dispense with mathematics in describing the properties of these objects, as David Malament points out, this would be "no victory at all! Even a generous nominalist like Field cannot feel entitled to quantify over *possible dynamical states*" (Malament 1982, p. 533).

Further objections to Field's nominalization program focus on his conservativeness claim, and more generally on his entitlement to the logical resources he makes use of. There is an ambiguity in the informal account of conservativeness given above, as it is not specified whether, by "consequence," we mean a semantic or deductive notion (the two notions coincide for first-order logic, but come apart when we move to second-order theories such as Field's nominalistic version of Newtonian gravitational theory). As Stewart Shapiro (1983) points out, and as Field himself acknowledges in *Science without Numbers*, the conservativeness claim fails for Field's second-order scientific theories, if by conservativeness we mean deductive conservativeness. Field therefore must make use of a semantic notion of conservativeness, one that, as Field puts it, involves "*consequence* rather than *provability*" (Field 1980, p. 115, n.30). But what is this logical consequence relation that, for second-order theories, outstrips provability in a formal system? It is standard to define semantic consequence in terms of truth in all models: *O* is a semantic consequence of theory *S* iff *O* is true in all models in which the axioms of *S* are true. But this set theoretic definition is clearly unavailable to Field if he wishes to reject the existence of mathematical objects such as sets. Field's solution is to adopt a primitive modal operator "it is logically possible that," in terms of which the logical consequence relation can be defined, thus rejecting the reduction of logical modalities to truths about the sets. Field remains queasy about second-order logic, stating in *Science without Numbers* that he shares "the feeling that the invocation of anything like a second-order consequence relation is distasteful" (p. 38). However, a restriction to first-order quantification would present significant difficulties for Field's project as outlined here (though Field has, in *Science without Numbers* and elsewhere presented partial defenses of a first-order nominalization of science). As with neo-logicism, it appears that the most attractive version of Field's program would require a defense of second-order logic.

In recent years, an alternative line of attack against the indispensability argument has been developed, in the light of observations from mathematical and scientific practice that present difficulties for Quine's confirmational holism. For example, Penelope Maddy (1992, 1997) points out that neither mathematicians nor scientists behave as though they see the existence of mathematical objects as a matter for empirical science to decide: mathematicians do not pay close attention to developments in empirical science in order to guide their attitude to, for example, the continuum hypothesis and its resolution; scientists do not take special care about the mathematics they make use of so as to minimize ontological commitments in line with their preference for ontological economy. Maddy also, crucially, notes that scientific practice appears to speak against confirmational holism given the presence of apparently

indispensable idealizations in our scientific theories (such as the simplifying assumption made in fluid dynamics that fluids are continuous substances). And, she notes, scientists often look for more than mere indispensability to our theories to confirm their belief in the existence of a theoretical object (witness the hunt for the Higgs Boson). All this suggests that indispensability to our best scientific theories may not be enough to provide evidence for the existence of an object of a particular sort.

Nominalists who wish to avoid the hard work of dispensing with mathematics in empirical science have seized on these observations as providing a way in for arguing that mathematical posits, though potentially indispensable, are not among the confirmed portions of our scientific theories. Focus has turned to the role played by mathematical posits in our scientific theories, with a new wave of mathematical fictionalists suggesting that such a role could be played by mere fictions as well as by really existing objects. In his defense of fictionalism, Mark Balaguer suggests that the role of mathematical posits in our scientific theories is to enable us to express an ultimately nominalistic content (the claim that the physical world is the way it would have to be for our scientific theories to be true). He holds that the supposed causal isolation of mathematical objects means that it is coherent to accept quantification over mathematical objects in our scientific theories while believing only the nominalistic content of those theories:

We can think of it this way: if all the objects in the mathematical realm suddenly *disappeared*, nothing would change in the physical world; thus, if empirical science is true right now, then its nominalistic content would remain true, even if the mathematical realm disappeared; but this suggests that if there never existed any mathematical objects to begin with, the nominalistic content of empirical science could nonetheless be true. (Balaguer 1998, p. 132)

Similarly, Jody Azzouni (2004), Joseph Melia (2000), Stephen Yablo (1998, 2005), and myself (2010) have also suggested that Platonism can be resisted if we view the role of mathematical posits in our scientific theories as primarily to enable us to express a nominalistic or physical content, since this is work that could be done by theoretical fictions.

The Quinean picture of mathematics as confirmed by its role in empirical science has been defended against these nominalistic proposals by Platonists such as Alan Baker (2005) and Mark Colyvan (2001, 2002, 2010). Debate has focused on the role played by mathematical posits in empirical science, with Baker and Colyvan arguing that mathematical posits do much of the same work as physical posits in our theories. In particular, they argue, the role

of mathematical posits is not simply to express a true nominalistic content; mathematical posits sometimes, they claim, play an indispensable *explanatory role*. In response to this, some nominalists (e.g. Melia 2002) argue that mathematical posits are never explanatory in this way, while others (e.g. Leng 2005) claim that the kind of explanatory role played by mathematics is also one that could be played successfully by mere fictions. Mark Colyvan (2010) argues that dealing with the issue of the presence of mathematics in explanatory contexts is likely to push those looking for an “easy road” to nominalism back toward Field’s hard road of dispensing with mathematics in the formulation of our scientific theories.

Whether or not Colyvan is right that Field’s project of rewriting our scientific theories so that they do not quantify over mathematical objects is essential for the defense of mathematical fictionalism, it is hard to see how a fictionalist view of mathematics could be defended without doing at least some of the hard work required by Field’s version of fictionalism. Insofar as fictionalists make use of a notion of “truth-in-the-story” for mathematics to underpin a sense in which the claims of standard mathematics are correct if not literally true, it is likely that this will be cashed out in terms of the notion of logical consequence (what is true-in-the-story of Peano arithmetic is presumably the axioms and their logical consequences). At the very least this will require a notion of logical consequence that is not reduced to its model-theoretic correlate, so like Field, such fictionalists appear pushed toward accepting primitive modality. And insofar as they wish the notion of “truth-in-the-story” to be complete for a given mathematical domain (so that, for example, for any arithmetic sentence *S*, either *S* is true-in-the-story of Peano arithmetic or not-*S* is), it is likely that they will have to accept second-order axiomatizations of mathematical theories, and therefore a second-order consequence relation.⁷ So “easy road” fictionalism does not avoid the hard work of justifying the logical assumptions behind the fictionalist view of mathematics, and in particular of showing that these assumptions are not themselves mathematical assumptions in sheep’s clothing.

Another option for defending nominalism in the light of the indispensability argument is to accept the truth of the mathematics used in science but to reject the inference from mathematical truths to mathematical objects, for example, by providing a nonstandard semantics for mathematical claims. This kind of approach is developed by Geoffrey Hellman (1989), in his modal-structural interpretation of mathematics, drawing on the following suggestion of Putnam (1967):

“Numbers exist”; but all this comes to, for mathematics anyway, is that
(1) ω -sequences are *possible* (mathematically speaking); and (2) there are

necessary truths of the form “if α is an ω -sequence, then . . .” (whether any concrete example of an ω -sequence exists or not). (p. 301)

In Hellman’s elaboration of this suggestion, we use the second-order Peano axioms to characterize the notion of an ω -sequence (natural number structure) categorically. Thus, if $PA^2(0, N, s)$ are the second-order Peano axioms with primitive terminology $0, N, s$ (with 0 a singular term, N a predicate symbol, and s a function symbol),⁸ the claim that an ω -sequence is possible can be expressed using a logical possibility operator “ \Diamond ” and by replacing $0, N$, and s by appropriate variables as:

$$\Diamond \exists x \exists X \exists f (PA^2(x, X, f)) \quad (1)$$

If $A(0, N, s)$ is a sentence in the language of PA , then the claim that A is true becomes the conjunction of (1) with the claim that, necessarily, if there is an ω -sequence then A is true when interpreted about that ω -sequence, or, in other words,

$$\Box \forall x \forall X \forall f (PA^2(x, X, f) \supset A(x, X, f)) \quad (2)$$

Where “ \Box ” is an “it is necessary that” operator, definable in terms of the “ \Diamond ” operator as “ \Box ” = “ $\neg \Diamond \neg$.”

Like Field, Hellman takes his “it is logically possible that” operator as a modal primitive (which he calls a “logico-mathematical modality” (p. 15)). He is also, clearly, in his use of second-order axiomatizations, just as committed to the use of this modal operator to characterize a second-order consequence relation. On the face of it, though, Hellman’s account looks like it should have an easier time than do fictionalist versions of mathematical nominalism in dealing with the indispensable presence of mathematical posits in our scientific theories. Yes, Hellman can claim, mathematical theories receive empirical confirmation, but

Rather than commitment to certain abstract objects receiving justification via their role in scientific practice, it is the claims of possibility of certain types of structures that are so justified. (Hellman 1989, p. 97)

The difficulty with this is that when mathematical language gets intertwined with the language of empirical science it is not at all straightforward to provide a modal-structural interpretation of these mixed mathematical/empirical claims that provides the right results. We do not, after all, wish to interpret our empirical theories as claiming merely that their (mixed mathematical/

empirical) laws are *logically possible* and their predictions follow logically from these laws. Instead, Hellman has to introduce a “non-interference proviso” so that the modalities work in the right way. In short,

we must *stipulate* from the outset that the only possibilities we entertain in employing the “ \Box ” are such as to leave the actual world entirely intact. (ibid., p. 99)

That is, the claims of our scientific theories are interpreted as hypothetical claims about what would be true were there a realm of objects entirely separate from the actual objects in the physical world, about which our mathematical theories could be interpreted as asserting truths. We may well wonder how it is that we are meant to draw conclusions about the actual world from these hypothetical claims. Presumably, we do so because the noninterference proviso requires that all truths about the actual world are held fixed, so if we infer that an entirely nonmathematical claim *would* be true *were there* noninterfering mathematical objects satisfying our mathematical assumptions, then it follows from the noninterference proviso that that claim is actually true. Hellman’s account here looks suspiciously close to the “easy road” versions of fictionalism: our mathematically stated scientific theories are to be trusted because we believe them to have things right in their picture of the concrete realm—or in other words, because we take it that their nominalistic content is true.

Putnam’s initial suggestion of a modal interpretation of mathematics, as taken up by Hellman, is presented as a direct response to Benacerraf’s (1965) discussion of multiple reductions. Indeed, Benacerraf himself suggests a structuralist resolution of the difficulty, stating that arithmetic is

The science that elaborates the abstract structure that all progressions have in common merely in virtue of being progressions. (1965, p. 291)

In Hellman’s modal structuralism, the “abstract structure” that all progressions have in common is eliminated with via the modal framework, but other noneliminative versions of structuralism view abstract structures as existing over and above any actual or possible instances. Thus Stewart Shapiro (1997) calls his Platonistic version of structuralism *ante rem* (as opposed to *in re*) structuralism, in parallel with the medieval distinction between *in re* and *ante rem* universals. Shapiro’s abstract structures exist *ante rem*, over and above any of their instances, whereas for Hellman, structures need only be thought of as existing when instantiated (*in re*).

Ante rem versions of structuralism, such as the accounts developed by Shapiro and Michael D. Resnik (1997), are versions of Platonism, and as such face Benacerraf’s knowledge problem for *abstracta*. Resnik adopts a

thoroughgoing Quinean holistic epistemology, drawing no lines between the claims of mathematics, logic, and empirical science. Shapiro takes a different tack, and offers a three-fold epistemology of structure. Some structures are known directly through their concrete instances, and perhaps from projecting the pattern seen in concrete instances to infinity. Other structures are grasped linguistically, through restrictions of rich languages to sublanguages where equivalence relations are treated as identities (e.g. starting from the language of Peano arithmetic we may choose to identify numbers based on their remainder when divided by seven, thus learning about the “addition and multiplication modulo 7” structure). But many mathematical structures cannot be known in either of these ways. Such structures are, according to Shapiro, known through our grasp of the logical coherence of their defining axioms. Shapiro hypothesizes (in his “coherence axiom” for structure theory), that “any coherent theory characterizes a structure, or class of structures” (Shapiro 1997, p. 95). “Coherence,” for Shapiro, maps on to what we have been calling “logical possibility”: a coherent collection of axioms is one that is satisfiable (i.e. has a set theoretic model) though, to avoid circularity in his account of structure existence, Shapiro notes that “Satisfiability is a model of coherence, not a definition of it” (ibid., p. 13). Shapiro thus accepts “coherence” as “a primitive, intuitive notion” (ibid., p. 135), and holds that it is via our grasp of the coherence (logical possibility) of second-order axioms (as implicit definitions) that we can come to know that they truly describe a structure (modulo his coherence axiom). *Ante rem* structuralism, at least in Shapiro’s version, thus appears to be in the same boat as neo-logicism, fictionalism, and modal structuralism in relation to its commitments to second-order logic and primitive modalities.

Unlike fictionalism and modal structuralism, *Ante rem* versions of structuralism are able to account for the claims of standard mathematics as truths using a standard semantics (with singular terms such as “the number 2” referring to positions in mathematical structures). As a result, one should expect them to have an easier ride than do fictionalists and modal structuralists when it comes to interpreting the “mixed” claims of empirical science, since these can be interpreted straightforwardly as truths without invoking any nonstandard interpretation of their mathematical components. Nevertheless, there are some difficulties with providing a straightforward account of these truths, given that structuralism as an account of mathematics is designed primarily to treat mathematical, rather than empirical, structures. For purely mathematical structures, the coherence axiom is relatively unproblematic—we take the mathematical realm to be plenitudinous, with abstract structures corresponding to every coherent collection of axioms. But our empirical scientific theories are expressed in the language of set theory with nonmathematical objects as ur-elements: we allow physical objects to be collected into sets, and posit the

existence of functions from these sets to real numbers (e.g. the mass function, whose domain is the set of all massive objects and whose range is the set of positive real numbers). It is not at all straightforward to see how sets of non-mathematical objects should fit into a structuralist ontology—they are mathematical objects (being sets), but are not *simply* the positions in structures. Furthermore, if we applied the coherence axiom to the laws of our empirical theories it would follow that these theories truly describe a mathematical structure, but the same would go for “empirical” theories that make wildly false claims about the physical world, so long as they are coherent. In order to distinguish between good scientific theories and bad ones it looks as though a structuralist will have to claim not just that they are true *of some structure or other*, but that they are true *when their physical terminology is interpreted as referring to actual physical objects*. But this looks rather close to the fictionalist’s claim to the *nominalistic adequacy* of our scientific theories: they are not just coherent, they are also true in their nominalistic content.

Shapiro’s coherence axiom is meant to bypass Benacerraf’s epistemological worry by showing how our beliefs could reliably reflect how things are with abstracta even in the absence of any contact with the abstract realm. If the coherence principle is correct, then the mathematical realm contains structures that satisfy the axioms of any coherent theory. So if our mathematical beliefs track coherence, then they will also track the truths about the mathematical realm even in the absence of any contact with that realm.⁹ A similar epistemological story is provided by Mark Balaguer (1998) in support of full-blooded (or plenitudinous) Platonism. Full-blooded Platonism (FBP) assumes that the mathematical realm is plenitudinous in the sense of having enough mathematical objects to provide models for any consistent mathematical theory. Again, *modulo* its existence assumption, FBP is able to explain the accuracy of our mathematical beliefs despite positing no contact between ourselves and mathematical objects; so long as our beliefs track the facts concerning consistency and logical consequence they will be true of some objects or other. And again, to avoid circularity in this account of the mathematical realm, Balaguer takes logical consistency to be an irreducibly modal notion. Balaguer hopes, however, to avoid reliance on second order logic by building extralinguistic elements into our mathematical theories: we have, he says, a full conception of natural numbers that includes axioms but also our pretheoretic intuitions about the kinds of objects that the theory is meant to pick out, and these are enough to pin down a unique (up-to-isomorphism) model for the natural numbers. Where our “full conception” is not powerful enough (as, for example, may be argued to be the case for our concept of set), Balaguer embraces nonuniqueness and allows that multiple nonisomorphic systems of objects may satisfy our mathematical theories. In this case, a supervaluationist truth-theory is given: in ZFC, since the continuum hypothesis (CH) is true in

some universes that satisfy our full-conception of the sets, and false in other such universes, CH lacks a determinate truth value.

As with *ante rem* structuralists, Balaguer is not explicit about how “mixed” mathematical objects, such as sets whose members are physical objects, fit into his “plenitudinous” picture of the mathematical realm, but again it is reasonable to assume that Balaguer takes there to be all the “mixed” mathematical objects there logically possibly could be, holding the empirical “facts” fixed. So again we have something like a noninterference proviso, with “good” empirical theories being those that are not just consistent, but correct in their nominalistic content.

Despite their ontological differences, then, at least in the contemporary philosophical accounts of mathematics surveyed here we see similar difficulties arising again and again. With the possible exception of neo-logicism, all these accounts have some difficulties with the intertwining of mathematical and empirical languages into the “mixed” facts emphasized by Quine. This intertwining speaks strongly in favor of a uniform semantics, as Benacerraf notes, but unless one bites the bullet to adopt a thoroughgoing Quinean holistic epistemology there will remain concerns about our ability to know the mathematical portions of these mixed claims, and attempts to bypass these concerns by suggesting that it is consistency or coherence that matters when it comes to matters of mathematical truth and existence inevitably raise the question again of what to make of these mixed truths where mere coherence is not enough. A natural line to take is that for mixed mathematical/empirical theories what matters is not just consistency, but consistency with the empirical facts, or, in other words, nominalistic adequacy, a notion that is in need of much clarification. Added to this, it appears that central to all the accounts considered is the assumption of a notion of logical possibility that is not reducible to deductive consistency or model-theoretic satisfiability. And all accounts go beyond first-order characterizations of mathematical theories, with most assuming that we can have a grasp of the second-order consequence relation despite the absence of a complete derivation system for second-order logic. Gödel’s first incompleteness theorem showed that (in second-order logic) there was more to the logical consequence relation than derivability in a formal system. In contemporary philosophy of mathematics it appears that we are still grappling with the implications of this result.

Notes

- 1 Steiner (1987) argues that Kant’s reading misrepresents Hume’s views.
- 2 Indeed, in Kant’s view (*Prolegomena* [4:273]), “The good company in which metaphysics would then have come to be situated would have secured it against the danger of scornful mistreatment; for the blows that were intended for the latter would

- have had to strike the former as well, which was not [Hume's] intention, and could not have been."
- 3 Its full title is worth repeating: "*THE ANALYST; OR, A DISCOURSE Addressed to an Infidel MATHEMATICIAN. WHEREIN It is examined whether the Object, Principles, and Inferences of the modern Analysis are more distinctly conceived, or more evidently deduced, than Religious Mysteries and Points of Faith.*" Berkeley's challenge is presented to mathematicians who use their supposed superior reason to reject religion. Having presented the obscurity of discussions of Newton's Fluxions or Leibniz's Differences, he asks, "But with what appearance of Reason shall any Man presume to say, that Mysteries may not be Objects of Faith, at the same time that he himself admits such obscure Mysteries to be the Object of Science?" (3).
 - 4 The banishment of geometry from arguments in analysis was an explicit goal of many involved in the "rigorization" process. "It is," Bolzano claimed in presenting his proof of the intermediate value theorem, "an intolerable offence against correct method to derive truths of pure (or general) mathematics (i.e. arithmetic, algebra, analysis) from considerations which belong to a merely applied (or special) part, namely geometry." Bolzano 1817, trans. Buss 1980, p. 160.
 - 5 It may still be wondered how a grasp of our mathematical *concepts* can guarantee the existence of *objects* falling under those concepts. In support of this picture, neo-logicians adopt a Fregean thesis about the primacy of language, viewing the existence of objects as falling out of the existence of meaningful sentences involving referring terms.
 - 6 Where "we say that concepts F and G *differ evenly* if the number of objects falling under F but not G, or under G but not F, is even (and finite)" (Boolos 1990, p. 215).
 - 7 Balaguer is happy to allow for incompleteness in the story of mathematics, so that there may be some truth-value gaps in some theories (such as, for example, set theory). He thinks second-order axiomatizations can be avoided even in places such as number theory where we may want to assert completeness, suggesting that our "full conception of natural number" may be enough to pin down all truths about the natural numbers even if there is no collection of first-order axioms that does so.
 - 8 In Hellman's version there is a single nonlogical primitive—the successor function. 0 is defined as the unique object that is not a successor of any object, and the axioms ensuring that the domain of quantification contains all and only the members of an ω -sequence. In this presentation I have kept 0 as a singular term and N as a predicate symbol as axioms expressed in these terms are more familiar.
 - 9 As such, this is an externalist epistemology for mathematics: we may in fact have mathematical knowledge in virtue of the truth of the coherence axiom, even if we do not (and cannot, on Benacerrafian grounds) know the truth of this axiom.

Bibliography

- Azzouni, J., 2004. *Deflating Existential Consequence: A Case for Nominalism*. OUP: Oxford.
- Baker, A., 2005. "Are there Genuine Mathematical Explanations of Physical Phenomena?" *Mind*, 114, pp. 223–38.
- Balaguer, M., 1998. *Platonism and Anti-Platonism in Mathematics*. OUP, Oxford.
- Benacerraf, P., 1973. "Mathematical Truth," *Journal of Philosophy*, 70, pp. 661–80. Reprinted in Benacerraf and Putnam 1983, pp. 403–20.
- , 1965. "What Numbers Could Not Be." *Philosophical Review*, 74, pp. 47–73. Reprinted in Benacerraf and Putnam, 1983, pp. 272–94.

- Benacerraf, P. and Putnam, H., eds, 1983. *Philosophy of Mathematics: Selected Readings*. 2nd edn. CUP: Cambridge.
- Berkeley, G., 1734. *The Analyst; Or, A Discourse Addressed to an Infidel Mathematician*. London: J. Tonson. Edited by David R. Wilkins 2002, and published online at: www.maths.tcd.ie/pub/HistMath/People/Berkeley/Analyst/Analyst.pdf
- Bolzano, B., 1817. *Rein analytischer Beweis des Lehrsatzes, dasszwischen je zwey Werthen, die ein entgegengesetztes Resultat gewähren, wenigstens eine reele Wurzel der Gleichung liege*, Wilhelm Engelmann. Translated by S. Russ as *Purely Analytic Proof of the Theorem that between Any Two Values which Give Results of Opposite Sign, there Lies At Least One Real Root of the Equation*, *Historia Mathematica* 7 (2), 1980, pp. 156–85.
- Boolos, G., 1998. *Logic, Logic, and Logic*. Cambridge, MA: Harvard University Press.
- , 1990. “The Standard of Equality of Numbers.” In G. Boolos, ed., *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge, MA: CUP. Reprinted in Boolos 1998, pp. 202–19.
- Colyvan, M., 2010. “There’s No Easy Road to Nominalism.” *Mind*, 119, pp. 285–306.
- , 2002. “Mathematics and Aesthetic Considerations in Science.” *Mind*, 111, pp. 69–74.
- , 2001. *The Indispensability of Mathematics*. Oxford: OUP.
- Ewald, W. B., ed., 1996. *From Kant to Hilbert: A Source Book in the Foundations of Mathematics*, 2 volumes. Oxford: OUP.
- Field, H., 1989. *Realism, Mathematics, and Modality*. Oxford: Blackwell.
- , 1985. “On Conservativeness and Incompleteness.” *Journal of Philosophy*, 81, pp. 239–60. Reprinted with a postscript in Field 1989, pp. 125–46.
- , 1980. *Science without Numbers: A Defence of Nominalism*. Princeton, NJ: Princeton University Press.
- Frege, G., 1902. “Letter to Russell.” In van Heijenoort 1967, pp. 126–8.
- , 1893/1903. *Grundgesetze der Arithmetik* H. Pohle, Jena. §§1–52. Translated and edited as *The Basic Laws of Arithmetic: Exposition of the System* by Montgomery Furth 1964, Berkeley, CA: University of California Press.
- , 1884. *Die Grundlagen der Arithmetik, eine logisch mathematische Untersuchung über der Begriff der Zahl*, W. Koebner, Bresleau. Translated by Austin, J. K., 1953 as *The Foundations of Arithmetic*, 2nd edn. Oxford: Blackwell.
- , 1879. *Begriffsschrift, eine der arithmetischen nachgebildete Formalsprache das reinen Denkens*, L. Nebert, Halle. Translated and edited by Bynum, T. W., 1972, *Conceptual Notation and Related Articles*. Oxford: OUP.
- Gödel, K., 1986. *Collected Works, Vol. 1: Publications 1929–1936*. Oxford: OUP.
- , 1947. “What is Cantor’s Continuum Problem?” Revised and expanded in Benacerraf and Putnam 1983, pp. 470–85.
- , 1931. On Formally Undecidable Propositions of *Principia Mathematica* and Related Systems I. Reprinted in Gödel 1986, pp. 144–95.
- Hale, B. and Wright, C., 2001. *The Reason’s Proper Study*. Oxford: Clarendon.
- Hellman, G., 1989. *Mathematics without Numbers*. Oxford: Clarendon.
- Hilbert, D. 1926. “On the Infinite.” Translated by Erna Putnam and Gerald J. Massey from *Mathematische Annalen*, 95. In Benacerraf and Putnam 1983, pp. 161–90.
- Hume, D. 1975. *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. Reprinted from 1777 edition, 3rd edn, L. A. Selby-Bigge, ed. Oxford: Clarendon.

- Kalderon, M., ed., 2005. *Fictionalism in Metaphysics*. Oxford: Clarendon.
- Kant, I., 1781. *Critique of Pure Reason*. Translated by Norman Kemp Smith 1929. London: Macmillan.
- , 1783. *Prolegomena to Any Future Metaphysics*. Translated and edited by Gary Hatfield (1997, revised edition 2004), Cambridge: CUP
- Kline, M., 1972. *Mathematical Thought from Ancient to Modern Times*. Vol. III. Oxford: OUP.
- Kneale, W. and Kneale, M., 1962. *The Development of Logic*. Oxford: Clarendon Press.
- Leng, M., 2010. *Mathematics and Reality*. Oxford: OUP.
- , 2005. "Mathematical Explanation." In C. Cellucci and D. Gillies (eds), *Mathematical Reasoning and Heuristics*. London: King's College Publishing, pp. 167–89.
- Maddy, P., 1997. *Naturalism in Mathematics*. Oxford: Clarendon.
- , 1992. "Indispensability and Practice." *The Journal of Philosophy*, 89, pp. 275–89.
- Malament, D., 1982. "Review of *Science without Numbers: A Defense of Nominalism*." *Journal of Philosophy*, 79, pp. 523–34.
- Melia, J., 2002. "Response to Colyvan." *Mind*, 111, pp. 75–9.
- , 2000. "Weaseling Away the Indispensability Argument." *Mind*, 109, pp. 455–79.
- Putnam, H., 1967. "Mathematics Without Foundations," *The Journal of Philosophy*, 64, pp. 5–22. Reprinted in Benacerraf and Putnam 1983. pp. 295–311.
- , 1979. *Philosophical Papers*, i. *Mathematics, Matter and Method*, 2nd edn. Cambridge: CUP.
- , 1975. "What is Mathematical Truth?" *Historia Mathematica*, 2, pp. 529–43. Reprinted in Putnam 1979, pp. 323–57.
- , 1971. *Philosophy of Logic*. New York: Harper and Row. Reprinted in Putnam 1979, pp. 323–57.
- Quine, W. V. O., 1980. *From a Logical Point of View*. 2nd edn. Cambridge, MA: Harvard University Press.
- , 1951. "Two Dogmas of Empiricism" *Philosophical Review*, 60, pp. 20–43. Revised version in Quine 1980, pp. 20–46.
- Resnik, M. D., 1997. *Mathematics as a Science of Patterns*. Oxford: OUP.
- Russell, B., 1902. "Letter to Frege." In van Heijenoort 1967, pp. 124–5.
- Shapiro, S. 1997. *Philosophy of Mathematics: Structure and Ontology*. Oxford: OUP.
- , 1983. "Conservativeness and Incompleteness." *Journal of Philosophy*, 80, pp. 521–31.
- Steiner, M., 1987. "Kant's Misrepresentations of Hume's Philosophy of Mathematics in the *Prolegomena*." *Hume Studies*, 13, pp. 400–10.
- van Heijenoort, J., 1967. *From Frege to Gödel: A Sourcebook in Mathematical Logic, 1879–1931*. Harvard University Press, Cambridge: MA.
- Whitehead, A. N., and Russell, B., 1910, 1912, 1913. *Principia Mathematica*, 3 vols, Cambridge: CUP. 2nd edn, 1925 Vol. 1, 1927 Vols 2, 3. Abridged as *Principia Mathematica to *56*, Cambridge: CUP, 1962.
- Yablo, S., 2005. "The Myth of the Seven." In Kalderon 2005, pp. 90–115.
- , 1998. "Does Ontology Rest on a Mistake?" *Aristotelian Society, Supplementary Volume*, 72, pp. 229–61.

17 Philosophy and Language

Barry C. Smith

1 Introduction

The capacity for language is unique to humans. Many other species have signaling systems by which they communicate but only human beings use language in an unbounded way to express their thoughts and feelings and to describe the world around them, both as it is, and as it might be. The importance of language to minded creatures like us, including its interpersonal use, its creative possibilities, its function of classifying and categorizing, has attracted philosophical attention, and led to large claims concerning the indispensable and inescapable role of language in the constitution of reality and our conceptual scheme. To some philosophers not only do we conduct philosophical inquiries in language but also our inquiries cannot concern anything but language. For these thinkers all intellectual life is a struggle with language. Such vast claims have been enthusiastically embraced by disciplines outside philosophy but have been cautiously resisted by analytic philosophers.

Philosophers of language in the analytic tradition have mainly focused their attention on two central concerns: the ability of language to express and communicate our thoughts; and the relation of language to reality. Broadly speaking, both issues bear on the language's representational powers: its ability to encode thought and portray aspects of reality. We shall explore both issues, as well as considering the relative priorities given to questions concerning language's connection to the mind and language's connection to reality. Is language subservient to thought, a mere outer clothing for inner states, with the contents of our thoughts relating the mind to reality? Or, on the contrary, is it because the mind apprehends language that it is able to exploit language's representational power to connect it to things beyond the immediately perceptible environment? The first important trichotomy is: thought/language/reality. The second trichotomy that concerns us, which we shall go on to explore, is that between syntax/semantics /pragmatics.

2 The Nature of Linguistic Phenomena

In pursuing the relations between language, mind, and world, it is necessary to be clear about the scope of the phenomena we are calling “language”; in particular it is important to distinguish language from communication. Communication can be nonverbal. A nod or a wink will do at times., By contrast, *linguistic* communication is far more expressively powerful. Just by using words we can talk about people, places, and things that are not present, and on any given day we will hear and use sentences that we have never heard or used before. Our capacity to produce and understand novel sentences stands in need of explanation, and it is worthy of note that we understand novel sentences just as easily as those with which we are familiar, suggesting that the same system is at work in understanding familiar and novel sentences. This system is a large part of what makes these capacities linguistic.

So what accounts for the distinctively *linguistic* character of linguistic communication? And what, precisely, is the nature of the linguistic phenomena under study? This is a fundamental question for the philosophy of language—but one that is seldom asked;¹ and rather than turning to empirical linguistics for an answer, philosophers often start by assuming that the facts to be explained are readily accessible as part of our surroundings. Speakers communicate using words and sentences to convey what they mean. So words and sentences, or rather the utterance of words and sentences, becomes the focus for philosophers wishing to account for our capacity to engage with linguistic meaning. For example, Michael Dummett tells us:

The philosopher . . . is, very properly, perplexed by the notion of meaning. He quite rightly regards it as an extraordinary thing, demanding explanation, that words—noises that issue from our mouths or marks we make on paper—should *have* meanings. (Dummett 1991, p. 14)

Notice the identification of words with noises or marks on paper as if these expressions were easily identified in advance of specifying their meanings. This is a dubious assumption. Of course, it is easy to pick out words in a language we understand, whose meanings we already know. However, we only need to try listening to speech in a language we do not understand to see how difficult it is to tell where one word ends and another begins, and this should remind us what an achievement it is to pick out word boundaries in the continuous sound signal of speech. Even in a familiar language the task is daunting “because the acoustic realizations of a given word can vary greatly depending on speech rate, speaker’s voice features, context, etc.” (Dehaene-Lambertz et al. 2005, p. 21). Identifying the words and the phonemes that constitute them, despite the variability of their articulation, is due to processes of

which we have no conscious awareness. These processes have to impose order and segmentation on the sound stream that confronts us in order for us to hear words as occurring in that sound signal. Thus, claims for the readily accessible nature of linguistic items have to be qualified. They are readily accessible to creatures with a certain internal organization and this must not be forgotten when thinking fundamentally about what languages are.

In order to treat acoustic signals as speech sounds we need to relate words and phonemes to acoustic signals. Beyond that we need to relate words to one another in sentences, since it is sentences that enable us to produce and comprehend indefinitely many thoughts. As Frege puts it:

It is remarkable what language can achieve. With a few sounds and combinations of sounds it is capable of expressing a huge number of thoughts, and in particular, thoughts which have not hitherto been grasped or expressed by any man. How can it achieve so much? By virtue of the fact that thoughts have parts out of which they are built up. And these parts, these building blocks, correspond to groups of sounds, out of which the sentence expressing the thought is built up, so that the construction of the sentence out of parts of a sentence corresponds to the construction of thoughts out of parts of a thought. And we may take a thought to be the sense of a sentences, so we may call a part of a thought the sense of that part of the sentence which corresponds to it. (From "Logic in Mathematics," in Frege 1979, p. 225)

Here, Frege talks of a sentence as built up out of a group of sounds, but strictly speaking, it is phonemes standing at one remove from sounds, and the words they compose, that make up sentences. And as Frege knew well, sentences are not just lists of words but crucial sites for the systematic relation between form and meaning. It is by having elements standing in relations of grammatical dependency to one another that the syntactic constituent structure of a sentence is determined: relations specified by the rules of a grammar. These syntactic rules play a crucial role in determining the groupings or arrangements, as well as possible rearrangements.

In syntax, "John loves Mary" fixes a relation between items that differ from "Mary loves John." Sentences are not reversible. We can say:

(1) John loves her

but not:

(2) *Her loves John

The grouping of elements that matters here is

[_{SNP} John [_{VP} loves _{NP} her]]

where

[__ loves her] is a constituent of the sentence, [John loves __] is not.

How we individuate languages depends on the way we articulate their syntactic structures. And how we should think of constituents and their syntactic properties, what such entities are, and where they are, is a big issue in the philosophy of linguistics.²

Philosophers, such as W. V. Quine, have tended to think of languages as sets of well-formed word-strings, where words are conventional sound-meaning pairs and, well-formedness is licensed by the grammar rules for the language in question.

This conception of language has been roundly criticized by Noam Chomsky, principally because grammars that only pronounce on whether strings of words are sentences will be descriptively inadequate as characterizations of any natural language. Chomsky's criticism of Quine's conception of a language, which he terms an E-language (where "E" stands for external, extensional, and extending beyond the speaker), helps to bring out why sentences are not merely word strings, and why their structures are not linear structures. The same linear string of words can express more than one thought depending on the way we parse its syntactic structure. Structurally ambiguous strings such as (3) can express at least two different thoughts.

(3) Peter saw the captain with the binoculars.

The prepositional phrase "with the binoculars" can either qualify the noun-phrase, "the captain" or the verb-phrase "saw," depending on whether it is the Captain or Peter who has the binoculars. Which sentence this word-string realizes depends on where the prepositional phrase belongs in the syntactic structure. If, along with Quine, we merely counted the number of well-formed word-strings there were in a language (i.e. those generated in accordance with the rules of the grammar) these would be fewer than the actual number of sentences in the language. Quine's proposed grammar only *weakly generates* the sentences of the language, while what is needed, according to Chomsky, is a grammar that *strongly generates* all the permissible structures of the language by assigning each of them a structural description, and for this we need a generative grammar.

According to generative grammar:

... the crucial properties of sentences are not revealed by thinking of them as they are outwardly presented to us, namely as strings of signs [or sounds] but rather by their unobservable grammatical structure ... What it is for something to be a sentence for a person is for it to be a grammatical structure that is apprehended and applied to certain perceptible objects [sounds or signs]. (Higginbotham 1991, pp. 555–6)

Such grammatical structures are remote from surface form and certainly not apprehended fully in consciousness. So instead of treating words and sentences as utterances of sounds, and treating languages as sets of behaviors, or social practices, Chomsky treats languages as internal to the minds of their speakers:

language has no objective existence apart from its mental representation. (Chomsky 1972, p. 169, fn 2)

As a result:

linguistic theory is mentalistic, since it is concerned with discovering a mental reality underlying actual behaviour. (Chomsky 1965, p. 4)

So it is not the sounds and signs people produce that constitute the subject matter of generative linguistics, but the *linguistic forms* people impose on those sounds and signs as a result of their internal states. It is only creatures like us, with the linguistic capacities we have, that can assign signs and sounds linguistic meaning and structure. On this conception of language as I-language (where “I” stands for internal, intensional, and idiolectal), the focus of linguistic inquiry shifts from the actual and potential behavior of speakers (of the sort that Quine and Dummett focus on) to the internal organization of individual speakers’ minds. The study of language is part of the study of mind, and so linguistics—the science of language—is treated as a branch of cognitive psychology. According to Chomsky, any account of language must be an account of a speaker’s knowledge of language:

The person who has acquired knowledge of a language has internalized a system of rules that relate sound and meaning in a particular way. The linguist constructing a grammar of a language is in effect proposing a hypothesis concerning this internalized system. (Chomsky 2006, p. 26)

It is worth noting that although the I-language theorist sees it to be internal to the mind of the speaker, this does not commit him or her to reducing language

to thought. The specialized part of the mind that enables human infants to acquire and develop language is a special purpose cognitive module, and questions about the relation between language and thought still remain to be settled. In fact, someone adopting an internal, cognitivist model of language could adopt a hard line in taking language to be not a vehicle for thought but a necessary condition for any kind of propositional thinking.

The competing conceptions of E-language and I-language both need to address and explain the same facts about a speaker's ability to produce and understand indefinitely many sentences. According to E-language conceptions, speakers succeed in communicating indefinitely many thoughts by taking advantage of an existing language in which words and sentences have meanings, so that what *sentences mean* in the language can explain what *speakers mean* in uttering them. (Famously, Paul Grice, proceeded in the other direction.)

The E-language theorist needs to say what it is for there to be a language with indefinitely meaningful expressions, and go on to explain what it is for speakers to make use of them. An account of language and meaning must be provided independently of speakers' *meaning something* if the former notions really can help to explain the latter. The language must be characterized first and then we must say something about what counts as *using* it. The biggest worry for the E-language theorist, is how to motivate an account of the syntactic structure of sentences and their recombinable parts—needed to explain the potential for a discrete infinity of sentences—without making any mention of speakers' psychology.

To make progress theorists of language of both E and I persuasion must say:

- (a) what a *language* is;
- (b) what *words* and *sentences* are;
- (c) what gives them *meaning*;
- (d) what counts as speakers' *uttering* or *using* them;
- (e) what speakers convey by their particular uses of words and sentences.

To do so, they will need, in addition to a syntax, a semantics and a pragmatics, and a principled way of relating them. Very roughly, semantic properties are those by which sentences succeed in representing the world, and syntactic properties are those by which strings of words succeed in being sentences. This already grants the theorist words, which are constituted by phonemes; so, strictly speaking, each theorist will need an account of the phonological properties that delimit the range of human speech sounds, and in addition an account in morphology of the word-like properties they determine. I shall say no more about phonology and morphology here, save to say that these create great difficulties for E-language theorists since there is no known way to align

phonemes with acoustic signals.³ Instead, I will concentrate on how syntax and semantics contribute to linguistic meaning and whether this is an existing system we can exploit to communicate *linguistically*. However, a full account of how language functions will have to explain how we get from the sounds or signs people make to what they mean.

3 The Naïve Semantic Picture

The same system seems to be at work in a speaker's understanding of familiar and novel sentences, and this hints at the possibility of commonalities at some underlying level that the semanticist searches for by attempting to construct a formal theory along the lines employed by logicians to uncover the logical forms of ordinary sentences. According to the Naïve Semantic Picture: what a speaker means on an occasion of utterance is what the words he utters on that occasion mean. His *meaning something* depends on his uttering *words* in a *particular grammatical order*; that is, one that constitutes a sentence. The meaning of the sentence as a whole depends on the meaning of its parts and the grammatical arrangements they stand in to one another.

In this way, an account of what *words and sentences mean*, together with an account of what it is for someone to utter particular words or sentences on an occasion, is supposed to provide an account of what it is for a *speaker to mean* what s/he does in uttering the words s/he utters on a given occasion. For such an account to work, we need an account of the meaning of words, the syntactic workings of the language, and the impact those syntactic workings have on sentence meaning.

Such a story requires a compositional theory of meaning: a theory of how the meaning of complex expressions depends on the meaning of their parts and the way they are put together. And in order to provide an illuminating account of sentence meaning without simply presupposing it, the theory of meaning was cast by Donald Davidson in terms of a theory of truth:

Speakers of a language can effectively determine the meaning or meanings of an arbitrary expression . . . and . . . it is the central task of the theory of meaning to show how this is possible. (Davidson 1984, p. 35)

A theory of truth shows how this is possible by providing a finite means for recursively specifying the conditions under each of the infinitely many declarative sentences of the language is true by seeing the truth-value of each sentence as determined by the truth-affecting properties of its parts and the ways they are combined syntactically. The meaning of each sentence amounts to its truth-condition: the way the world has to be for that sentence to be true. If a

sentence S says P, and P is the case, then sentence S is true because the world is as the sentence says it is. So if we specify that the sentence S is true if and only if P, we can take it that S says that P.

For Davidson, a truth theory for a language L should deliver every instance of the schema:

(T) S is true in L iff p

where “S” is replaced by a declarative sentence of L and “p” is replaced by some sentence in the meta-language with the same truth-value. The truth-theory is situated within a large theory of radical interpretation. To confirm the truth theory the interpreter needs to find out under what conditions speakers take their sentences to be true, by starting with a sample of sentences and working out how they come to have the truth conditions speakers take them to have on the basis of their distinguishable parts and their syntactic combination. He then uses the assumption of system to generate new sentences and assign them truth conditions and test whether speakers of the language hold the new sentences true under the conditions he has assigned to them in the truth theory. The aim is to find out which assignments of structure to sentences and semantic values to their parts could have resulted in the distribution of truth-values the radical interpreter finds among the sentences in question. In this way, the radical interpreter comes up with a theory of interpretation for the speaker’s utterances.

The machinery of the theory comes into play to prove T-sentences for each declarative sentence of the language from form finite axioms of the form:

- (R1) The reference of [_{NP} Clinton] = Clinton
- (R2) The reference of [_{NP} Obama] = Obama
- (S1) [_S[_{NP} ∂] [_{VP} smokes]] is true iff ref [_{NP} ∂] smokes
- (S2) [_S[_{NP} ∂] [_{VP} is [tall]]] is true iff ref [_{NP} ∂] is tall

Thus T-sentences for the L-sentences “Clinton smokes” and “Obama is tall” parsed as [_S[_{NP} Clinton] [_{VP} smokes]] and [_S[_{NP} Obama] [_{VP} is [tall]]] can be proved from the above axioms: for example,

- (a) “Clinton smokes”
- (b) “Clinton smokes” = [_S[_{NP} Clinton] [_{VP} smokes]]
- (c) [_S[_{NP} Clinton] [_{VP} smokes]] is true iff ref [_{NP} Clinton] smokes
- (d) The reference of [_{NP} Clinton] = Clinton
- (e) [_S[_{NP} Clinton] [_{VP} smokes]] is true iff Clinton smokes
- (f) “Clinton smokes” is true iff Clinton smokes

An alternative to Davidsonian truth theories would be a theory that told us what *proposition* a sentence expresses. If a sentence S expresses the proposition P, and P is true, then S is true. Truth-theory dispenses with talk of propositions, but they are useful to appeal to for expository purposes. For example, corresponding to

[_S[_{NP} Clinton] [_{VP} smokes]] and [_S[_{NP} Obama]] [_{VP} is [tall]]

we have:

<Clinton, being a smoker> <Obama, tallness>

where each semantically relevant constituent of the sentence corresponds to an element of the proposition.

Any semantic theory of this kind is going to give *compositionality* a leading role to play in explaining speakers' linguistic abilities and what it is for them to say something meaningful by uttering a string of words in their language:

If (but only if) speakers of a language can understand certain sentences they have not previously encountered, as a result of acquaintance with their parts, the semanticist must state how the meanings of those sentences is a function of the meanings of the parts. (Evans, 1985, p. 344)

It is by providing a systematic account of how the semantic properties of complex properties depends on the semantic properties of their parts that a compositional theory aims to explain how speaker and hearers are able to understand one another's utterances.

The question is whether an account of sentence meaning together with an account of the systematic workings of the language can carry the burden of explaining how speakers uttering strings of words succeed in meaning something. A potential source of trouble is that there is not always such a nice matching of the semantically relevant constituents of the sentence and the entities of the proposition.

4 Uttering Sentences

The naïve semantic picture takes it that the thought a speaker expresses by uttering a sentence on an occasion is determined by the literal meaning of the sentence uttered, where the literal meaning of the sentence is determined compositionally by the meaning of the words it contains and the way they are put together syntactically.

But what is it to *utter a sentence*? Here, we return to question (d), which we need to answer before we rely on the answer to question (c). And as we will see, doing so immediately returns us to questions (b) and (a). Is it sentences we utter or just the sounds that phonetically realize particular words and phonemes? As we have seen, sounds are not specific enough to fix semantics. Besides, two languages may share the same sounds but use them to articulate different words and sentences. Even within a single language there are homophones like “Check” and “Czech.” Let us grant for the sake of argument that the speaker produced a sequence of expressions. Which sentence was uttered?

As stated, there are two important qualifications to the idea that speakers utter sentences: (a) ambiguity, and (b) context-sensitivity.

(a) Grammatical strings can be ambiguous as a result of lexical ambiguity, as in (4) below, or structural ambiguity, as in (5):

- (4) Bills like this one cause trouble.
- (5) Tex likes exciting sheep.⁴

Since each of (4) and (5) can have more than one meaning we need to advert to the speaker’s *intentions* to know what s/he meant in uttering either of these word strings. Even if we take different voicings of (4) as uttering the same sentence, whether we take “Bill” to be an act of parliament or a demand for payment, without reference to the speaker’s intentions on the occasion we cannot know what s/he meant. And in the case of (5) we need to know the speaker’s intentions to know which sentence s/he uttered.

(b) As well as words with stable, repeatable meanings, there will be sentences containing words whose reference changes from one occasion of use to another. These context-sensitive expressions are indexicals like “I,” “now,” “here,” “today.”

These expressions will have their reference assigned as a result of features of the context of utterance such as the identity of the speaker, the time, and the place. So although there is context-sensitivity about what is said by uttering a sentence in a given context, the role of context is fixed by the conventional rules that specify the meanings of the indexical expressions. However, there may be other, less obvious context-sensitive expressions: “local,” “foreigner,” “tall,” “rich,” whose meanings depend on less obvious or stable features of context. This gives rise to a further question: what are the limits on context-sensitivity?

A further problem is how we should think about sentences and their tokening in utterances. It may not be enough to talk about a sequence of uttered expressions because of phenomena like verb phrase ellipsis:

- (6) Lee loves his cat and so does Bill.

Is it Lee's cat or his own cat that Bill loves? We need to advert to hidden or phonetically null constituents of the syntactic structure of the sentence uttered [love his cat] to get at either interpretation and here we need to postulate a level of underlying linguistic representation at which the ambiguous string is made clear. Or else we need to posit just one syntactic structure as the one the speaker tokened. Linguists call this the level of *logical form* (or LF) for short—a level of syntactic representation remote from surface form. As we have seen, generative linguists take it to be mentally represented in the mind of the speaker and hearer.

Even if we have managed to sort out what is uttered—a sentence with hidden structure, for example—we still need to know what “elements” occur in such a structure. If there are grammatical constituents of a sentence that can be assigned *semantic* values then we must regard them as part of the expression uttered even if they are not pronounced, that is, phonetically null. The need to postulate an unpronounced “his” that can be assigned either to Lee or Bill as its reference is the only way to bring out the semantic ambiguity of an utterance of (6). But such a story requires us to give a rather radical answer to (b) and as a result an answer to (a) that is hard to square with most E-language accounts.

We want to know which sentence is uttered when a speaker produces an acoustic signal. But the trouble is the term *sentence* has a number of uses that should be distinguished. A useful classification of things we can call “sentences” is provided by Robert Stainton as follows:

(Sent) Three senses of “sentence”

- a. sentence syntactic: an expression with a certain kind of structure/form
- b. sentence semantic: an expression with a certain kind of content/
meaning
- c. sentence pragmatic: an expression with a certain kind of use.⁵

We can conceive of an utterance of (7a) below, for example, as consisting of a 4-tuple of the form in (7b), where the first member of the 4-tuple is the phonological representation *P*, the second the syntactic string Σ , the third the semantic content *M* with its logical form, and the fourth the speech act content, *C*.

(7a) Peter left.

(7b) </peter left/, [S [NP Peter] [VP left]], <left(peter)>, J asserts in C
<left(peter)>

Which of these elements we pay attention to in accounting for what a speaker uttered depends on our theoretical interests. But what we know is that syntax has a meaning-determining or meaning-shaping role.

As another example of how meanings and form are related we can consider the syntactic constraints on referential interpretation as shown by the indices indicating same or different referential interpretations:

John₁ shaved him₂
John₁ shaved himself₁,
John's₁ mother likes him₂
Mary₁ expected to feed herself₁,
I wonder who₁ Mary expected to feed herself₂

These facts about the referential dependence of one item on another, and of which interpretations we can and cannot give to these sentences, depends on purely structural configurations (whether some items “c-command” others).⁶

There is a significant difference between the surface form of a sentence and its underlying logical form. For example, there are often moved or displaced elements in a string, and the sentence structure is richer than the string that gets pronounced or inscribed:

John seems to be happy / John₁ seems t₁ to be happy.
John seems to Bill to want to leave / John₁ seems to Bill t₁ to want to leave.
John seems to Bill to doubt himself / John₁ seems to Bill t₁ to doubt himself₁.

Some movements of elements in the linguistic form of the sentence make no difference to the meaning of the sentence, for example, so-called quantifier floating:

All the children were in the garden / The children were all in the garden
There are also *scope* considerations:

(a) Everyone in the class speaks two languages.

Either everyone has their two languages, or there are two languages and everyone speaks them, depending on whether the quantifier “everyone in the class” takes wider scope over “two languages” or vice versa.

(b) A person from every city complained about it.

Here, “a person” takes narrower scope and “every city” takes wide scope over it.

On the Naïve Semantic Picture, there may be further facts about the speech act performed by uttering a given sentence (including what the speaker meant

in uttering it) but the meaning of the utterance will be due to the literal meaning of words and sentences determined by the rules of the language. Sometimes we are interested in properties of the utterance *as well as* the semantic properties of the sentence uttered. Sometimes it is just the latter. Speaking in a funny voice may convey more than meaning.

5 From Uttering to Meaning

Both *uttering* and *meaning* something are occasion-specific, one-off events. This makes them ideally suited to pair up. A speaker's uttering certain sounds, and his meaning what he does by doing so, is something that occurs at a particular time and place: a dateable, unrepeatable event. How are they related: How can uttering a sentence amount to meaning something? If we mean by a sentence something at least as detailed as the syntactic string Σ , then it will contain words as constituents.

But *words* and *sentences* are not temporally or spatially bounded, nor occasion-specific, provided at least that we accept a distinction between what words mean, and what a speaker means in uttering them.

Linguistic knowledge provides a speaker with the words and means to articulate and recognize indefinitely many sentences. So what relations can be traced between the linguistic means available to speakers through their standing knowledge and the things they mean on occasion by their utterances? How should we connect one with the other? This is a daunting task for philosophers, linguists, and psychologists. What else is involved and what other resources can be drawn upon in accounting for how uttering becomes, and is recognized as, a case of meaning something?

Much remains to be done, and much of the work is empirical. What relations are there between the first and second elements in (7b)? Are there intermediate levels between them that are theoretically significant, and/or psychologically real? Answers to these questions may give us greater confidence in postulating the third and fourth elements in (7b).

The intermediaries or levels posited between uttering and meaning belong to a theory of language, and the postulated properties or levels may be subject to different interpretation by rival theorists who may conceive these levels, their roles and relations quite differently. The postulated items and the relations between them bear on the distinctions between syntax, semantics, and pragmatics. Some will privilege syntax and semantics, others syntax and pragmatics.

The distinction between semantics and pragmatics is the subject of a large controversy, and it should be said at the outset that perhaps there is no fully

satisfactory unifying account of the relations between uttering and meaning, nor any single theory that can take care of all the examples. This is quite likely. Extremes include uttering without meaning anything at all (babble), and meaning without uttering; for example, when you did not utter a sound you may have meant something by your silence. Nevertheless, we should still be able to figure out the indispensable constraints on the mapping between uttering (sound) and meaning.

6 Saying and Meaning

So far, we have been supposing that what speakers mean is what the sentences they utter mean, where this is a matter of the meanings of their parts and the way they are put together syntactically. Is the naïve semantic picture *roughly* right? There is a danger of losing the one-off feel of meaning something by using the timeless meaning of words. Do what words and sentences mean on occasions of use depend on their timeless meaning, or the *speaker's meaning* what s/he does by them on that occasion? We do not want to say that the speaker utters something and this is what *it* means. Compare:

- (a) what a word or sentence means (almost timeless?)
- (b) what a speaker means on an occasion (act-bound)

Perhaps there is a need for an intermediary notion or level: *what is said*—by uttering a sentence with that meaning on a particular occasion of use. Is this a semantic or pragmatic notion?

This is a controversial issue: compare *what is said by a speaker* in uttering a sentence on a given occasion versus *what the sentence says*: as determined by meanings of the constituents of the sentence and how they are put together. There is work this distinction can do in distinguishing between *what is said* and *what (else) is meant or implied*. Sometimes the content conveyed by one speaker to another seems to go beyond, or depart from, what is literally expressed by the sentence. Thus, a gap seems to open up between the meanings of the words and sentences speakers utter and what they mean in uttering them. This gap provides one way of drawing the distinction between semantics—the study of the meanings of simple and complex expressions—and pragmatics—the study of how people use language for communicative purposes. “Semantics deals with *the literal meaning of words and sentences* as determined by the rules of the language, while pragmatics deals with *what users of the language mean* by their utterances of words or sentences” (Recanati 2010, p. 1). This is not the only way to draw the semantics-pragmatics distinction, as we shall soon see, but it has played a large part in the twentieth-

century tradition of semantic theorizing about language. It would allow us to say the following:

(S&P1) At times a speaker means *more* than what the sentence uttered literally means, but s/he means at least what the sentence uttered literally means.

7 Grice's Distinction between Semantics and Pragmatics

According to Paul Grice the semantics-pragmatics distinction corresponds to the distinction between what is literally said, or expressed, by an utterance of a sentence and what else is additionally communicated. It is the distinction between *what is said* and *what is meant*, by uttering a sentence in a given context. But here we need to take care to distinguish Grice's more technical notion of *what is said*—as fixed by the literal meaning of the sentence—from the ordinary, intuitive notion. Intuitively, we may talk about whether people say what they mean and mean what they say. This saying indicates that people can sometimes say one thing and mean another—this would not be permissible on Grice's understanding of the divide since people always mean what is said by uttering the sentence—that is, it is always part of what they assert—but they may mean or assert more besides:

(8) I'm tired⁷

The sentence in (8) can be uttered when I am asked if I want to go to the cinema, where clearly what is meant is that I do not want to go to the cinema this evening. For Grice, I have asserted that I am tired. That is, what is said by uttering (8), and what is meant over and above what is said depends on a *conversational implicature* that the hearer must infer from what was said and the background information. The speaker can always cancel the implicature—"I didn't say I didn't want to go to the cinema—all I said was that I was tired. That's why it's not an implication of what I said." Grice distinguished the pragmatics of "but" and "and," and also "and" and "and then" but regarded them as all having the same semantics. "She was poor but she was honest" differs from "She was poor and she was honest"; "She shouted at her boss and was fired," "She was fired and she shouted at her boss" are different, but only pragmatically, according to Grice. Imagine someone saying to a colleague who asks where his wife is:

(9) Either your wife is in the Library or she's with her lover.

What is said is true if your wife is in the Library. What the utterance suggests is that she has a lover, but it does not literally *say* that. According to Grice, we typically mean what we say but we may mean more. What is meant or *implicated* over and above what is said (determined on the basis of the literal meanings of the parts of the sentence and their syntactic arrangement) is cancellable. What is meant is inferred from what is said (construed as the literal meaning of the sentence).

For Grice, what is said by an utterance of a declarative sentence is fixed entirely by the literal meaning of the sentence: what is said must correspond to “the elements of [the uttered sentence], their order and their syntactic character” (see chapter 5 of his 1989 book).

We often think, as Kent Bach notes, that “the constituents of what is said must correspond to the constituents of the utterance” (Bach 1994). But is this the case? Let us follow Bach and call this the *Syntactic Correlation Thesis*:

(SCT) Every element of the proposition expressed—what is said by uttering S—must be the value of a constituent of S.

The corollary holds too: every semantically relevant part of the sentence must contribute an element to the proposition expressed.

Now consider some of Kent Bach’s examples:

- (10) You won’t die [from that cut]
- (11) I haven’t eaten [today]
- (12) I’ve nothing to wear [to the party]

What is meant (as shown by the additional material in square brackets) is meant *instead* of what is said (in Grice’s sense—the literal meaning of the sentence), and not in *addition* to what is said. In (10), the speaker is not saying to the child that it is immortal *and* saying that it will not die from the cut. It is therefore more tempting to suppose that *what is said* by the utterance of these sentences is something nonliteral. That is, the literal meaning of the sentence uttered does not fix what is said. But then what does?

Speakers can communicate all sorts of things by their utterances: their tone of voice can convey how angry they are, they can be sarcastic or ironical, they can speak knowing someone is overhearing the conversation, they can hint at something not said in so many words. Thus there is often a huge gap between the meanings of the words and sentences uttered by a speaker and what a speaker means in uttering them. Drawing the semantics-pragmatics distinction as Grice did prescinds from all of these difficulties but we may wonder whether these additional features enter into not just into what is meant, but also, pace Grice, into what is said.

So far we have seen that one way to articulate the semantics-pragmatics distinction is to distinguish the linguistic information encoded by expressions and the nonlinguistic information brought into play as a result of using those expressions in a context. But now the question arises as to whether the linguistically encoded information in a declarative sentence always has to determine a proposition.

Kent Bach's makes a semantics-pragmatics distinction as follows:

(S) Semantic information is information encoded in what is uttered—these are stable linguistic features of the sentence—which together with any extralinguistic information provides (semantic) values to context-sensitive expressions in what is uttered.

(P) Pragmatic information is (extralinguistic) information that arises from an actual act of utterance . . . pragmatic information is generated by, or at least made relevant by, the act of uttering. (Bach 2001)

Right from the start notice that Bach has to appeal to extralinguistic information about the circumstances of utterance in his account of semantic information in order to talk about how context-sensitive expressions like "I," "now," "here," and "that" get their semantic properties (references) assigned.

Is the semantic information Bach talks about enough to show how a sentence expresses a proposition in a context of utterance? The answer is no. And since Bach does not require the speaker to mean what is said according to the literal meaning of the expressions in the sentence and the way they are put together he is free to reject Grice's distinction between semantics and pragmatics.

(S&P2) At times the speaker means something *other* than what the sentence literally means.

On Grice's picture, every element of *what is said* corresponds to some element of the sentence uttered (including empty categories, syntactic ellipsis, etc.) and is determined by their syntactic arrangement. On Bach's alternative semantic-pragmatic picture of how meanings of words and sentences and speaker's meaning are related, the semantic content of a sentence typically underdetermines what is said. Sentence (linguistic structure) does not determine who or what the speaker is referring to: sentences may contain demonstrative or indexical elements. "I'll be there tomorrow." However, contextual effects on what is said are not restricted to parameters of time, place, or utterer.

Facts about speaker's meaning, and pragmatic processes are needed to derive truth conditional content of an utterance. This is the view that context affects what we mean by "red" or by "tall." If pragmatics is needed to fix the

values of these expressions then the semantic content of whole sentences containing these expressions depends not just on the semantics of these expressions and how they are arranged syntactically, but also on the pragmatics of these expressions: what speakers mean by their use of these expressions in certain contexts, or what hearers can work out on the basis of pragmatic processes and what the content makes available. A speaker will have failed to communicate successfully and convey what she is saying if neither she nor the context provide enough information from the hearer to augment what is encoded in the linguistic material uttered.

8 Beyond Grice's Distinction

There is not always such a nice matching of the semantically relevant constituents of the sentence and the entities of the proposition. What happens when the content of the proposition seems to go beyond what is literally expressed by the sentence? How should we account for the mismatch between what is said and the literal meaning of the sentence? Take Bach's example:

(10) You won't die

Said to a child, what is literally expressed is that he is immortal and it is highly unlikely that this is what is said, if what is said is also meant. According to Bach, what is meant—that he will not die from that cut—is something meant *instead* of what is said (in Grice's sense), not in *addition* to what is said. It is therefore more natural to suppose that what is said by the utterance of (10) is something nonliteral, as in (11):

(11) I haven't eaten

(12) I've nothing to wear

I do not say I have never eaten anything, or in (12) that I have no clothes. Grice himself worried about the case of metaphor and irony and claimed that in using sentences in this way speakers only "make as if to say" what such sentences express. This is a concession to the difficulty with his notion of what is said. Notice that none of the natural readings of (10) to (12) can easily be supposed to be implicatures—something inferred from what is said.

The problem arises for Grice because what is said must correspond to "the elements of [the uttered sentence], their order and their syntactic character." But as Bach points out, this is not always the case. Consider:

(13) Everyone laughed

(14) Tony is ready

- (15) Britney has arrived
- (16) It's raining

According to Bach (1994) these sentences do not yet express a proposition and need to be completed somehow, or have a missing element filled in in the context in which they are used:

- (17) Everyone[in the class] laughed
- (18) Tony is ready [for . . . ? to . . . ?]
- (19) Britney has arrived [at the party]
- (20) It's raining [here]

According to Bach the literal meaning of such sentences express what he calls a *proposition radical*, something that needs completion to produce a full proposition. For this reason Bach says that such sentences are semantically underdetermined. However, when they are uttered in a context, hearers have no difficulty interpreting the speaker and recovering a propositional content, something that can be true or false. Somehow, the full propositional meaning is *implicit* in the use of such sentences, and we rely on this fact in conversing with one another. People seldom state part of what they intend to get across or precisely state exactly what they mean. We speak loosely and leave elements out while relying on our hearer to be able to understand or work out what we are saying. We do it all the time. (As I just did then.)

So Bach denies that Grice's distinction between what is said and what is implicated (meant or communicated in addition) exhausts the notions of meaning that surrounds the utterance of sentences. For Bach, we need a three-fold distinction between:

What is said (literally expressed) / what is implicit in (what is said) / what is implicated.

For Bach, the key notion in uttering a sentence is what is *implicit* in the utterance of it. This revision to Grice's account allows Bach to say that when we use a sentence nonliterally as in (13)–(16), we do not mean what is said, but rather something else instead. This corresponds to the intuitive notion of saying one thing but meaning another. What we mean is the nonliteral meaning: for example, in (10), the speaker means that the addressee will not die of that cut. There could still be other meanings conveyed in addition just as Grice supposed, as a matter of conversational implicature. The utterance may carry the conversational implicature that the hearer should shut up and stop complaining about the cut. (Cancellable of course.)

Resistance to Bach's idea of an incomplete propositional element expressed by a sentence comes from those who would say each of (17) to (20) express minimal propositions, which are, perhaps admittedly, less than the speaker means, but which are nonetheless genuine and complete propositions in their own right. But is this objection correct? What are the relevant minimal propositions? In the case of (17) the minimal proposition might be of this form:

(17') Everyone (in the universe) laughed.

And what are the minimal propositions expressed by (18), (19), and (20)? Perhaps the following:

(18') Tony is ready (for something)

(19') Britney has arrived (somewhere)

(20') It is raining (somewhere in the universe)

Notice first that we need the explicitly stated extra material to state these minimal propositions. But could we say the sentences with the added constituents mean the same as the sentences without them? This does not seem right, since the negations of (17') to (20') do not appear to be equivalent to the negations of (17) to (20). When we say:

(21) Britney hasn't arrived.

this surely is not equivalent to the false and almost unintelligible:

(22) Britney hasn't arrived anywhere.

Besides, syntactically, we do not know how to fill out the minimal proposition corresponding to (18). Should it be Tony is ready (–to F), or is it (–for . . .)?

Bach does consider cases where a minimal proposition is expressed and where what is implicit in the utterance is an expansion of the proposition expressed, as in:

I haven't anything to wear [to the party]

I haven't eaten [today]

You won't die [of that cut]

I haven't eaten [lunch]

So there are two ways in which pragmatics is required to arrive at the implicit content of an utterance: in the case of sentences that *semantically underdetermine* the proposition communicated we resort to the process of

completion. In the case of sentences that express *minimal propositions* there is a need to resort to *expansion*. Both are pragmatic processes required to recover what the speaker meant or what was implicit in his or her utterance of the sentence. The proposition communicated is an expanded version of the one explicitly expressed—what is said by an utterance of the sentence. So, most of the time for Bach, what is communicated is an expansion or completion of what is said.

9 A More Radical Revision to Grice's Picture

However, not everyone agrees with Bach that what is expanded or completed is an expansion or completion of what is said. It is possible to maintain Grice's insistence that there is an exhaustive distinction between what is said and what is implicated but argue that this distinction does not line up with a distinction between what is explicit or implicit, or with a semantic/pragmatic distinction.

Francois Recanati, for instance (along with Carston, Sperber, Wilson, and others), insists that what is expanded and implicated in the examples immediately above is part of what is said. For Recanati, what is said is not identical to what is determined by the meaning of words in a sentence and its linguistic form, but rather is something pragmatically determined by what is said by the utterance in context. So what is implicit in the utterance of sentences in context is *part* of what is said. What is said, on this model, incorporates elements that do not appear in or correspond to elements of the sentence uttered. There will be parts of what is said that do not correspond to any element in the utterance (or even in the logical form of the sentence). These elements of the proposition expressed are what Recanati following Perry (1986) calls "unarticulated constituents." Examples of these constituents are represented by the completing or expanded elements we supply in thought—the elements represented in the square brackets above.

An "unarticulated constituent" is not part of the explicit sentence structure, nor part of the underlying logical form. It is not linguistically constrained but introduced and governed instead by pragmatic principles. The idea was first alluded to by John Perry in his 1986 article, "Thought Without Representation." When we say:

(23) It's raining.

We usually mean:

(24) It's raining [here].

But according to Perry we do not need to suppose the “hidden indexical” element is represented in the logical form of the sentence.

However, if the proposition expressed by a sentence is determined by elements, some of which are not in the logical form of the sentences, then we cannot suppose the meaning of a sentence is compositional, that is, determined by the semantic values of the parts and their syntactic arrangement (logical form). Semantics now seems deprived of its task and tools. Truth conditional semantics no longer concern natural language sentences if the meanings of the latter are noncompositional. The domain of truth conditions may be how a sentence is interpreted as a result of pragmatic processes. The semantics/pragmatics divide collapses and pragmatics claims the victory.

Some philosophers—such as Jason Stanley—have argued that context can make a contribution to the truth conditions of the sentence uttered just so long as the contribution goes via logical form. In other words, pragmatically figured out references always contribute to the truth conditions of what is uttered by assigning a value to a constituent in logical form. So long as there is a variable (perhaps hidden) in logical form that can be assigned a value that is determined pragmatically in context, we can insist that the semantics-pragmatics distinction is a principled one, and that it is semantics that dictates what the truth conditions of a sentence uttered in context will depend on.

Other philosophers—such as Francois Recanati—will insist that context can make a contribution to the truth conditions of an utterance without going via something semantically determined or specified, and so there are no truth conditions without pragmatic processing. In contrast to truth conditional semantics, Recanati calls this *truth conditional pragmatics*.

But before we examine this dispute more closely, let us return to Bach’s notion of semantic information (which we encountered in Section 7). Is it sufficient to allow us to assign truth conditions to the sentence uttered in a given context? Not always, it seems.

The sentence uttered in context may contain less semantic information (including extralinguistic information relating to context-sensitive expressions) than we need to fix truth conditions, as in (25), (26), and (27).

- (25) Steel is strong enough.
- (26) Joan went to the edge of the cliff and jumped.
- (27) He opened the door with a key.

In each case we need more information added to determine the proposition conveyed, that is, to determine what is true or false. Steel is strong enough for what? Jumped over, jumped back, jumped up? Opened by turning the key in the lock?

Bach holds that the semantic information encoded by the sentence uttered does determine *what is said*, but he does not claim that this is enough to determine the *truth conditions* of the utterance of (23)–(27) or the proposition conveyed in context by their utterance. Thus he breaks the link between the following intuitively related notions:

Literal meaning of sentence uttered = what is said = the proposition expressed = the truth conditions of the utterance.

For Bach, as we have seen, *what is said* is something like a proposition fragment or proposition radical as he calls it. It is something that needs completion by the hearer adding some nonlinguistic element in thought.

What is said + thought constituent(s) = implicature / plus pragmatic inference = what is meant

Robyn Carston believes similarly that the linguistic expressions uttered almost always underdetermine the truth conditions of the utterance or the proposition that their utterance conveys. She calls the resulting enrichment of the linguistically encoded information the *semantic explicature*, in contrast to Grice's conversational implicatures that we can derive over and above the literal meaning of the sentence.

Notice this way of drawing the semantics pragmatics distinction leaves rooms for Grice's distinction between proposition expressed and what else is conveyed, but it departs from Grice in supposing that the sentence uttered does not fix the proposition expressed. Pragmatics intrudes into fixing this and thus into fixing the truth conditions of the utterance.

For Bach, what is said is fixed by what is linguistically encoded but that does not fix the proposition expressed nor therefore the truth conditions of the utterance. For Carston, like Recanati, the notion of what is said goes with truth conditions and proposition expressed but is a pragmatically determined, not a semantically determined, notion. Carston and Recanati are truth conditional pragmaticists.

Returning to the issue of the extent to which the truth conditions of a sentence uttered are semantically specified, the key issue is exemplified by the very simple weather case we encountered earlier:

(16) It's raining.

When uttered in a context this sentence is usually interpreted as expressing the proposition that it is raining in the context of utterance. But how does the

sentence get these truth conditions assigned when there is no explicit mention of location?

Some have suggested that there is a hidden indexical or hidden variable in the logical form of weather predicates so that (16) should really be represented at the level of LF as:

(16') It's raining [at x]

where the default value for "x" is "here," the place of utterance. Is there any evidence that there is a hidden variable in the underlying syntactic structure of the sentence in (16)? The pragmatic story could be that there is no variable but that we can pragmatically add an unarticulated constituent by a purely pragmatic process of *free enrichment*. This is the suggestion of Francois Recanati.

Syntactic evidence could be found by showing that there is a variable that can be bound by a quantifier as in:

Wherever I go, it's raining

Clearly this means: [where I go (x)] it rains at x. It clearly does not mean: wherever I go it is raining here at a place I am at when I utter the sentence. So we have both a quantified reading—wherever I go it is raining—and a specific reading of the self-standing "It's raining" meaning it is raining here.

Must we accept the syntactic postulation of the extra variable position for weather predicates? Not necessarily. The question is whether, in addition to these two readings we can have an *indefinite reading* without a location specified. Recanati thinks we can, as in his weatherman example. People in a control room waiting for news of rainfall anywhere, who have sensors all over the globe, will say "It's raining" with the indefinite reading as soon as any sensor alarm rings. The information conveyed is not about a specific location. If he is right about the example then there is no hidden variable in the syntax. For Recanati this means that when we have a specific reading we have to add an unarticulated constituent by a pragmatic process of free enrichment to get the truth conditions and what is said by an utterance of "It's raining."

Recanati rejects other interpretations of his example such as it is raining somewhere or its raining on earth. The first would have to show there was a hidden quantifier—hardly an uncontroversial way to show there was a hidden variable reading. The second reading would have to mean it was raining across the earth, according to Recanati, on the model of its raining in London. But that is not what is said. This is less convincing.

10 Contextualism

According to contextualists, sentences never express complete propositions independently of context, however explicit a speaker is. The content of what is said is *always* underdetermined by the linguistic material uttered.

For Recanati, what is said—the content of the utterance—is fixed by pragmatic processes together with what the linguistic material uttered provides. There are *primary pragmatic processes* and *secondary pragmatic processes*. The former are mandatory and unconscious inferences computed when we recover the content of the utterance. These are processes that determine what is said by a speaker who utters a sentence in a context. The secondary pragmatic processes are conscious, reflective inferences that allow us to derive what the speaker meant in saying what he or she said. These processes are used to work out the conversational implicature in Grice's sense, of what is said.

So like Bach, Recanati has a distinction between *what is said* and *what is meant*, but again not drawn as Grice draws the distinction (or as Bach does).

Recanati's primary pragmatic processes have three ingredients. First, *saturation*: the assignment of values to content-sensitive expressions such as indexicals and demonstratives. Second, *modulation*, whereby the extension of a term can be altered by the sentential context, that is, the meanings of other words in the sentence:

- (29) Jean eats rabbit
- (30) Jean wears rabbit
- (31) Tom's car is red [not all over]
- (32) The ball is red [all over]

Third, *free enrichment*, where constituents are added to fix the truth conditions of the utterance:

- (33) Jack and Jill went up the hill [together]
- (34) John and Mary are married [to each their]
- (35) Joan went to the edge of the cliff and jumped [over]

11 Against Contextualism

Stanley argues that the process of free enrichment will overgenerate spurious readings of the sentence meaning. If the added material—the unarticulated constituents—are not linguistic then why can't we add anything we like, without linguistic constraint? But we can't. We can't freely enrich (36) as follows:

- (36) John loves Mary [sixpence]

But why can we not enrich to suggest a monetary scale for love? The answer would be that are only two argument places for “loves” and we cannot just put in what we want. So free enrichment is not free but linguistically constrained.

Stanley looks for cases where we know precisely what to add and argues these must be linguistically specified and so are not candidates for free enrichment.

(37) All the students in his class passed the exam [for students in that class]

For Stanley and Szabo (2000), for each quantified noun-phrase there must be a function from context of utterance to some domain restriction within the class of things of kind N, which picks out the relevant domain of quantification.

This is the so-called problem of incomplete definite descriptions, where we talk about “the baby,” “the class,” “the room,” “the fridge,” etc. without suggesting there is only and only one baby, class, room, or fridge in the world. Is there a semantic or only a pragmatic solution to fixing which object uniquely satisfies the description?

12 Semantic Minimalism

Cappelen and Lepore (2006) argue that for sentences not containing context-sensitive expressions from the Basic Set articulated by David Kaplan: I, now, here, today, yesterday, etc.,

(Whatever the context) the sentence “S” expresses the proposition that S.

Thus “John is ready” is true iff John is ready and nothing more. We have to decide on the circumstances for evaluating the sentence, but we have specified a minimal proposition.

To establish these minimal propositions as the semantic content of an utterance as opposed to the speech act content, they use several tests of context sensitivity:

- (a) Intercontextual indirect speech disquotational reports;
- (b) Collective reporting across different contexts.

In (a) cases, I can quote you as saying that John is ready, though I cannot use the very same sentence you use if you deploy an indexical. In (b) cases, if Tony says “John is ready” and you say “John is ready,” I can say, Tony said

that too. Again, this marks a contrast with the use of pure indexicals like “I” and “now.”

Are these tests conclusive? It is not clear. Some people would accept collective reporting with indexicals (see Stojanovic 2009). Does moderate contextualism lead to radical contextualism? The arguments are (as yet) inconclusive.

13 Nonsentential Assertions

Robert Stainton argues that just by uttering a subsentential phrase we can make assertions that can be true or false:

- (38) Nice dress
- (39) Fifth floor, please
- (40) French
- (41) From Brazil

Someone could utter (38) when seeing a woman pass by. (39) would be said getting into an elevator. Seeing someone puzzled by the aroma of a cigarette I could helpfully say (40), and when contemplating someone’s dazzling display I could say to an onlooker (41). Can semanticists reject these cases? What should they say? There is reason to resist the idea that they are just cases of ellipsis. And once again the problem is deciding what was uttered. The work to get from what speakers utter in context to what they mean remains challenging and will require a good deal of collaboration between linguists and philosophers. It is well under way.

Notes

- 1 The topic has been more widely covered in the philosophy of linguistics. See Devitt 2006, 2008, Smith 2006, 2008, Collins 2007, 2008, Pietroski 2008, and Ludlow 2009.
- 2 The symbols S, NP and VP represent grammatical categories of sentence, noun phrase and verb phrase. For more, see Devitt 2006, Smith 2006a, b, Collins 2008, and Ludlow 2009, 2011.
- 3 For more on the philosophical significance of speech sounds see Smith 2009 and Bromberger and Lasnik 2012.
- 4 I owe this example to Michael Devitt.
- 5 Stainton 2006, p. 31.
- 6 In terms of the tree-geometry by which we represented hierarchical relations between syntactic items, an expression α c-commands an expression β when the first branching node dominating α dominates β . For more on c-command and binding see Chomsky 1995.
- 7 This nice example is Dan Sperber’s.

References

- Bach, K., 2005. "Context *Ex Machina*." In Z. Szabo, ed., *Semantics vs Pragmatics*, pp. 16–44. Oxford: OUP.
- , 2001. "You Don't Say." In *Synthese*. Kluwer Academic Publishing. 128, pp. 15–44.
- , 2000. "A Reply to Stanley and Szabo." *Mind and Language*. 15, pp. 262–83.
- , 1994. "Conversational Implicature." *Mind and Language*. 9, pp. 124–62.
- Bezuidenhout, A., 1997. "Pragmatics, Semantic Underdetermination and the Referential/Attributive Distinction." *Mind*. 106 (423), pp. 375–409.
- Cappelen, H. and Lepore, E., 2005. "Radical and Moderate Pragmatics: Does Meaning Determine Truth Conditions?" In Z. Szabo, ed., *Semantics vs Pragmatics* Oxford, OUP, pp. 111–64.
- , 2004. *Insensitive Semantics*, Malden MA: Blackwell, pp. 1–59, 143ff.
- Carston, R., 2002. "Linguistic Meaning, Communicated Meaning and Cognitive Pragmatics." *Mind and Language*. 17, pp. 127–48.
- , 1988. "Implicature, Explicature and Truth-theoretic Semantics." In R. Kempson, ed., *Mental Representations*. Cambridge: CUP; Also in Davies 1991.
- Chomsky, N., 2006. *Language and Mind*. Cambridge: CUP.
- , 2005. *The Minimalist Program*. Cambridge, MA: MIT Press.
- , 1972. *Language and Mind* (Harcourt, Brace, Johanovich).
- , 1965. *Aspect of the Theory of Syntax*. Cambridge, MA: MIT Press.
- Collins, J., 2008. "Knowledge of Language Redux." *Croatian Journal of Philosophy*. 22, pp. 3–43.
- , 2007. "Review of *Ignorance of Language*." *Mind*, 116 (462), pp. 416–23.
- Dehaene-Lambertz, G., Pallier, C., Serniiclaes, W., Sprenger-Charolles, L., Jobert, A., and Dehaene, S., 2005. "Neural Correlates of Switching from Auditory to Speech Perception." *NeuroImage*, 24, pp. 21–33.
- Davidson, D. 1984. *Inquiries into Truth and Interpretation*, Oxford: Oxford University Press.
- Davies, S., 1991. *Pragmatics: A Reader*. New York, Oxford: OUP, Blackwell.
- Devitt, M., 2008. "Explanation and Reality in Linguistics." *Croatian Journal of Philosophy* 23, pp. 203–31.
- , 2006. *Ignorance of Language*. Oxford: Clarendon Press.
- Dummett, M., 1991. *Frege: Philosophy of Mathematics*. London: Duckworth.
- Elugardo, R. and Stainton, R., 2004. "Shorthand, Syntactic Ellipsis, and the Pragmatic Determinants of What Said." *Mind and Language*, 19 (4), pp. 442–71.
- Evans, G., 1985. *Collected Papers*. Oxford: OUP.
- Frege, G., 1979. *Posthumous Writings*. Translated by P. Long and R. White. Oxford: Basil Blackwell.
- Grice, P., 1989. "Logic and Conversation." In P. Grice, ed., *Studies in the Way of Words*. Cambridge, MA: Harvard University Press, pp. 1–40.
- Higginbotham, J., 1991. "Remarks on the Metaphysics of Linguistics." *Linguistics and Philosophy* 14 (4), pp. 555–66. Reprinted in Carlos Otero (ed.), *Noam Chomsky: Critical Assessments*, Routledge, London, 1994.
- Ludlow, P., 2009. Review of Devitt 2006a. *Philosophical Review*, 118 (3), pp. 393–402.
- , 2011. *Philosophy of Generative Linguistics*. Oxford: Oxford University Press.

- Neale, S., 2002. "On Being Explicit: Comments on Stanley and Szabo." *Mind and Language*, 15 (2–3), pp. 284–94.
- , 1993. *Descriptions*, Ch. 3 "Context and Content." Cambridge, MA: MIT Press.
- Origg, G. and Sperber, D. 2000. "Evolution, Communication and the Proper Function of Language." In P. Carruthers and A. Chamberlain, *Evolution and the Human Mind*. Cambridge: CUP, pp. 140–70.
- Perry, J., 1986. "Thought Without Representation." *Proc. Aristotelian Soc. Supp.* Vol. 60, pp. 137–52.
- Pietroski, P., 2008. "Think of the Children." *Australasian Journal of Philosophy*, 86 (4), pp. 657–69.
- Recanati, F., 2010. *Truth-conditional Pragmatics*. Oxford: OUP.
- , 2005. "Literalism and Contextualism: Some Varieties." In G. Preyer and G. Peters, eds, *Contextualism in Philosophy: Knowledge, Meaning, and Truth*. Oxford: OUP, pp. 171–96.
- , 2003. *Literal Meaning*. Cambridge: CUP.
- , 2001. "What is Said." In *Synthese*. 128, (1–2), pp. 75–91.
- , 1989. "The Pragmatics of What is Said." *Mind and Language*. 4 (4), pp. 295–329.
- Schiffer, S., 1995. "Descriptions, Indexicals and Belief Reports: Some Dilemmas." *Mind*. 104 (413), pp. 107–31.
- Smith, B., 2009. "Speech Sounds and the Direct Meeting of Minds." In C. O'Callaghan and M. Nudds, *Sounds and Perception: New Essays*. Oxford: OUP, pp. 183–210.
- , 2008. "What Remains of Our Knowledge of Language?" *Croatian Journal of Philosophy*, 8 (22), pp. 557–75.
- , 2006b. "Why We Still Need Knowledge of Language." *Croatian Journal of Philosophy*, 6 (18), pp. 431–57.
- , 2006a. "What I Know When I know a Language." In E. Lepore and B. Smith, eds, *The Oxford Handbook to Philosophy of Language*, Oxford: OUP.
- Sperber, D. and Wilson, D., 1986. *Relevance: Communication and Cognition*. Oxford: Blackwell.
- Stainton, R., 2006. *Words and Thoughts*. Oxford: OUP, chs 1–3.
- , 2005. "In Defence of Non-Sentential Assertion." In Z. Szabo, ed., *Semantics vs. Pragmatics*. Oxford: OUP, pp. 383–457.
- , 1995. "Non-Sentential Assertions and Semantic Ellipsis." *Linguistics and Philosophy*. 18 (3), pp. 281–96.
- Stanley, J., 2005. "Semantics in Context." In G. Preyer and G. Peters, eds, *Contextualism in Philosophy*. Oxford: OUP, 221–53.
- , 2000. "Context and Logical Form." *Linguistics and Philosophy*. 23 (4), pp. 391–434.
- Stanley, J. and King, J., 2005. "Semantics, Pragmatics, and the Role of Semantic Content." In Z. Szabo, ed., *Semantics vs. Pragmatics*. Oxford: OUP, pp. 111–64.
- Stanley, J. and Szabo, Z., 2000. "On Quantifier Domain Restriction." *Mind and Language*, 15 (2–3), pp. 219–61.
- Stojanovic, I., 2009. "Semantic Content." *Manuscript* 32, 123–52.
- Travis, C., 1997. "Pragmatics." In B. Hale and C. Wright, eds, *A Companion to The Philosophy of Language*, Oxford: Blackwell, pp. 87–106.
- Wilson, D. and Sperber, D., 2002. "Pragmatics, Modularity and Mind-reading." *Mind and Language*, 17 (1–2), pp. 3–23.

18 Meaning, Normativity, and Naturalism

Richard Gaskin

1 Kripke's Skeptical Paradox

An outstanding event in contemporary philosophy of language was the publication in 1982 of Saul Kripke's *Wittgenstein on Rules and Private Language*. In this book Kripke argued for the startling thesis that there is no fact of the matter about what we mean by our words and, in general, no fact about what words mean. There is no fact constituting my meaning one thing by a word rather than something else. Ascriptions of meaning are nonfactual; they do not have truth-conditions; they are neither true nor false. The argument appears to proceed epistemologically. Kripke maintains that, even under ideal conditions of epistemic access—even under conditions in which I have epistemic access to all relevant facts—I can find nothing that could *justify* my claim to mean one thing rather than another. From there he moves to the metaphysical conclusion that nothing *constitutes* my meaning one thing rather than another. But in reality there is no transition: that the argument is already metaphysical, not epistemological, is guaranteed by the appeal to ideal conditions.¹ Note that though I here, following Kripke, make use of a first-person perspective in setting up the problem, that is an adventitious feature of the scenario described; the puzzle has nothing essentially to do with the question “how is first-person authority for intentional states possible?,”² and can be recast in purely second- or third-person terms. It concerns how meaning, not how anyone's access to meaning, is possible.

Suppose I have performed only finitely many computations in the past (that will be true) and suppose that I have never performed an addition on numbers larger than 56. So $68 + 57$ is a calculation I have never yet performed. I carry out the computation and duly get 125 as my answer. Kripke asks: what *justifies* my claim to have reached the correct answer—that is, the answer that accords with what I mean, and have always meant, by “+”—and what *constitutes* 125 as the correct answer? A skeptic might claim that the correct answer is actually something else, say 5. This skeptic might say that, in the past, when I used

the symbol “+,” what I really meant was a function that we might denote by *, defined as follows:

$$\begin{aligned} x * y &= x + y, \text{ if } x, y < 57 \\ &= 5 \text{ otherwise.} \end{aligned}$$

What justifies my assertion that I did *not* mean * by “+” in the past, and what justifies our claim that the correct answer to $68 + 57$ is *not* given by $68 * 57$? Kripke considers several kinds of fact that might constitute my having in the past meant addition by “+” rather than the * or any other function, and rejects them all.

First, he considers facts about my previous nonlinguistic behavior. These, he claims, do not determine what I meant by “+.” For, by hypothesis, all my previous computations using the sign “+” dealt with numbers less than 57, so my previous behavior is compatible with the assumption that I meant *. Secondly, we have facts about previous explicit instructions I gave myself or was given by others. I was taught the *addition* rule, you might say—that is, the rule that yields 125 as the answer to $68 + 57$ —not the * rule. But all the instructions I gave myself or was given by others were communicated in words, and by hypothesis the examples I had all concerned numbers less than 57. What is to stop a skeptic offering a deviant interpretation of those words, so that what I learnt was the * rule, not addition? You might object: I was taught to go on *in the same way* in new cases, and to give the answer 5 to $68 + 57$ is not to go on in the same way. But if the rule I actually learnt was the * function, rather than addition, then to give that answer would indeed be what “going on in the same way” demanded. Explicit instructions, as Wittgenstein has made familiar,³ do not determine their own interpretation: they are capable of different interpretations. Thirdly, we might adduce facts about my mental history. Here Kripke considers two possible candidates.

One such candidate would be a mental image. But mental images do not determine meaning, because they, like explicit verbal instructions, stand in need of interpretation, and are capable of being variously interpreted. It follows that having a mental image when I think of a word is not *sufficient* for associating a definite meaning with that word. Nor is it *necessary*. We understand many words without entertaining any mental images, and even for words (e.g. “table”) where one might be tempted to appeal to a mental image as constituting that word’s meaning, as the Lockean tradition does, it is not necessary to entertain any image in order to count as understanding the word. This, again, is a point on which Wittgenstein repeatedly insists in the *Philosophical Investigations*. A further possible candidate that Kripke considers is a sui generis mental state, a special sort of mental state, one without an introspectible phenomenology, but whose obtaining constitutes my having meant addition

by “+” in the past. Kripke responds to this suggestion by saying that in a sense it is irrefutable, but that it leaves the nature of meaning wholly mysterious. How could our finite minds embrace such an infinitistic state?

In connection with the first suggestion for a meaning-constituting fact, namely past nonlinguistic behavior, Kripke also considers a proposal that focuses not on past *overt* behavior but on past *dispositions* to behave. When I learnt the rule for addition I acquired a *disposition* to perform indefinitely many new calculations in certain ways. Perhaps my having acquired a particular disposition justifies my now returning the answer 125 to $68 + 57$? Kripke rejects the suggestion on the basis that it confuses the descriptive with the normative. A disposition determines causally what I *will* do; it does not determine what I *ought* to do. To suppose that it determines what I ought to do is to collapse competence and actual performance. Kripke gives two symptoms of the dispositional account’s failure to yield a kind of fact that could constitute my having in the past meant addition by “+” rather than the * (or any other) function.⁴ In the first place, I may be a competent adder even though I make mistakes—and have the disposition to make mistakes—in certain circumstances. For example, if I try to perform a very large addition under unfavorable conditions (let us suppose that I am very tired), I may make, and be disposed to make, a mistake. But our willingness to say this shows that we do not simply identify the function that I intend to compute with whatever I actually end up doing. I intend to *add*; but the answer I actually give in the imagined scenario is not the *sum* of the very large input numbers, but some other deviant, star-like function of them. So what I intend to do—what I understand by “+”—cannot simply be identified with whatever function I actually end up computing. In the second place, dispositions are finite, but the function that I internalize when I learn the meaning of “+” is infinitistic in its scope. Some numbers are so large that I would die before I computed their sum—so I have no disposition to give whatever is their sum as the answer—but we nevertheless want to allow that, in attempting to compute $m + n$, where m and n are such enormous numbers, I am attempting to compute their *sum*, and not some nonstandard function that maps two numbers to their sum if they are sufficiently small but fails to map them to any number at all if they are sufficiently large.

Kripke has argued that none of the above candidates succeeds in providing a fact that can both *justify* my assertion that I *meant* addition by “+” in the past and *constitute* my having formerly meant addition by “+.” But if that is right, then surely nothing *now* can justify my confidence that I *mean* addition by “+,” or can constitute that fact. This for two reasons. First, a similar argument to the one we have gone through can be run in the future concerning my current behavior and mental states.⁵ But secondly (and this is perhaps closer to what Kripke had in mind: see 1982, p. 11), the point is that, if there *were* something that constituted my having meant addition by “+” in the past,

then I could, now, simply appeal to that in justification of the way I go on. For example, I might simply *remember* how I meant “+” in the past, and decide to go on now *in the same way*. But if it is not determinate what I meant by “+” in the past, then I cannot appeal to it now in justification of my present and future practice. Finally, the argument plainly generalizes to all words and all speakers. In setting up the paradox, we made the assumption that we knew what we were talking about, that *our* use of the word “addition” referred to addition and not to a star-like function. But that provisional supposition must now be relinquished. Meaning evanesces. We reach, so Kripke concludes, the “incredible and self-defeating conclusion, that all language is meaningless” (1982, p. 71).

2 Straight and Skeptical Solutions

We may distinguish between a “straight solution” to the paradox, which meets it head-on and tries to refute it by producing a meaning-constituting fact, and a “skeptical solution,” which accepts the skepticism, and tries to accommodate it by arguing that it is not as intellectually devastating as it initially appears. Kripke offers a solution of the latter sort. He agrees with the skeptic that ascriptions of meaning are not fact-stating and do not have truth-conditions. Rather, they have assertibility-conditions: by implication, assertibility-conditions should replace truth-conditions as the central notion in the construction of a systematic theory of meaning. The claim that I mean addition by “+” will be *assertible* just if my practice agrees with the rest of the community’s practice in using the sign “+.” The community will be prepared to adopt me as a member and accept that I know the rule for “+” just if my responses to questions of the form “What is $m + n$?” agree with theirs. A statement like “Mary means addition by “+” “ is not to be regarded as fact-stating, or as having truth-conditions, but is to be interpreted as a gesture of acceptance, a mark of the community’s willingness to treat Mary as a member of the linguistic club: she can be relied on to manipulate the sign “+” in the same way as the other club members.

The skeptical solution has elicited many objections from commentators, and I think it is fair to say that it is generally regarded as unsatisfactory. In my view, there are essentially three problems with it. The first takes its cue from a natural worry to the effect that the skeptical solution is really a straight solution in disguise. For surely it offers a factual, truth-conditional account of what meaning ascriptions consist in—just a different one from the realist’s story. It tells us that Mary’s meaning addition by “+” *consists in* her agreeing with the community’s practice, but this contention is surely tantamount to the claim that it is *true*—that it is a *fact*—that Mary means addition

by “+” just if she agrees with the relevant community responses. And the fact that her responses so agree is what *justifies* her (and our) claiming that she means addition by “+.” So there apparently is a fact of the matter about meaning ascriptions after all. To deal with this objection, Kripke would have to apply his nonfactualism not merely to meaning ascriptions, but also to claims concerning the agreement or disagreement of an individual’s practice with the relevant community’s practice, and that would seem to tow in its wake a quite general nonfactualism, one that was no longer limited to meaning ascriptions. The position would end up being one not of skepticism specifically about the factual nature of *meaning*, but of skepticism about *factuality* in general.⁶ Meaning would be in the same boat as everything else. But then, surely, the nonfactualism of the position simply undoes itself, for on a global nonfactualism there seems no appreciable difference between saying that there are no facts about anything, just (actual or hypothetical) practices of agreement or disagreement—the apparent factuality of these practices itself, presumably, being constituted by further agreement practices, and so on (a regress unfolds)⁷—and saying that there indeed is such a thing as factuality, as we always thought, and that the facts are constituted by communal decision.

But are facts to be settled by communal decision? Can the community not go wrong? This is the second problem with Kripke’s skeptical solution.⁸ We have provided a standard of assessment for the individual’s meaning something: the individual counts as meaning addition by “+” just if his or her practice matches the community’s practice with this sign. But we have provided no standard of assessment for the community as a whole: there is no wider community for it to be measured against, so nothing seems to constitute the community’s meaning anything determinate by “+.” If, perhaps under the influence of some universally administered drug—this scenario reminds one of the opening of John Wyndham’s *The Day of the Triffids*, but there a few people escaped the meteorologically induced mass blindness, whereas here we are to suppose that everyone takes the drug—the entire linguistic community woke up tomorrow treating the sum of hitherto unencountered numbers in some deviant, star-like manner, then that would be the right answer, on Kripke’s approach. That is, we can make no sense of a communal mistake. As Wright puts it:

it is a community of assent which supplies the essential background against which alone it makes sense to think of individuals’ responses as correct or incorrect . . . None of us can unilaterally make sense of correct employment of language save by reference to the authority of communal assent on the matter; and for the community itself there is no authority, so no standard to meet. (1980, pp. 219–20)

In other words, Kripke's skeptical solution applies the dispositional analysis at the level of the community. That seems counterintuitive. Of course, if meaning is use—and there must be some sense in which this Wittgensteinian slogan is right—it no doubt follows that, at least in the long run, should the community continue to use “+” to denote a star-like function, then this deviant function is what that sign will in due course come to mean. But it only *comes to* mean that; it does not mean it immediately. At the point when the community practice changes, it is (a) true that its practice *does* indeed change, (b) true that the community is not now using “+” in conformity with its past meaning, and (c) true that, initially at least, the new usage is *incorrect*. How quickly the new usage *becomes correct*—as, if it is sufficiently widespread and consistently adhered to, it no doubt will—is immaterial: the point, which Kripke's skeptical solution cannot accommodate, is that it does not *start* by being correct.

This latter point is, evidently, a somewhat delicate one. Note three things about it. First, I have described the above scenario, as involving a “change” of community practice, from a perspective external to that community itself. The community in question might deny that any such change had occurred; it might assert that it had always meant “+” in the same way as the way in which it was now using that sign—that is, in a way that, viewed from my external perspective, seems to deviate from previous usage. And of course it *could* be the case that the community was right about that. But equally—and this suffices to refute Kripke's skeptical solution—it could be the case that the community was wrong.

Secondly, in this context talk of correctness and incorrectness is to be understood as mere shorthand for statements about truth and falsity. What changes when, as I said above, the practice of the community changes, is not the *meaning* of “+”—precisely not. What changes is that the community now misunderstands its own language, and accordingly gets the truth-values of relevant sentences wrong. The talk of “old” and “new” usage has to be understood in this sense: if by “usage” we intend something tantamount to *meaning*, then usage does not (immediately) change in the envisaged scenario; what changes is the penumbra of community practice with the relevant words, and in particular what the community takes to be true, and what false. It is in that sense that the new usage is (initially) incorrect. This is a point to which I shall return when I discuss normativity in more detail below. Thirdly, I have agreed that meaning is use—at least, it is *long-term* use—and what that slogan, in the present context, amounts to is the point that meaning is *communal* (long-term) use. Rule-following practices depend on brute, contingent agreements, on a communally shared “form of life.”⁹

That does not mean that rule-following can be *reduced* to those contingent agreements, conceived as describable in terms that do not presuppose notions of meaning, understanding, or in general of a rule. To think that would be, as

John McDowell says, to “locate ‘bedrock’ lower than it is,”¹⁰ that is, to suppose that intentionality can be grounded in phenomena that are not essentially intentional in nature. The metaphor of bedrock is taken from *Philosophical Investigations*:

“How am I able to follow a rule?” —If this is not a question about causes, then it is about the justification for my acting in *this* way in complying with the rule. Once I have exhausted the justifications, I have reached bedrock, and my spade is turned. Then I am inclined to say: “This is simply what I do.” (2009, I, §217)

Justifications, like explanations (compare I, §1), have to come to an end. But at the point where they do so, we do not have mere movements, or matchings of movements, but rule-governed behavior: “following according to the rule is FUNDAMENTAL to our language-game.”¹¹ We should “refuse to countenance sub-‘bedrock’ (meaning-free) characterizations of what meaning something by one’s words consists in.”¹² Note that the last sentence of the above quotation — “Then I am inclined to say . . .” — does not imply that “This is simply what I do” is necessarily the *right* thing to say. Wittgenstein merely records an inclination: whether it is correct or not depends on what exactly is meant. If “This is simply what I do” is supposed to open the door to sub-bedrock characterizations of my fundamental behavior, then the temptation to say it should be resisted. The action that, as Wittgenstein elsewhere tells us, “lies at the bottom of the language-game” (1989b, §204), is suffused with rule.

The third objection to Kripke’s skeptical solution concerns specifically the attempt to make assertibility-conditions supersede or do duty for truth-conditions. This move is simply incoherent. Anything that is asserted is asserted *as true*: if you assert that *p*, you assert that *it is true that p*. Truth is, in that sense, an empty predicate (“It is true that . . .” adds nothing to the *content* of what it is prefixed to). Hence if you assert that *p*, you cannot avoid asserting something with truth-conditions: at the very minimum you assert something that is true if and only if *p*. So there is no sense in which appeal to assertibility-conditions can *replace* appeal to truth-conditions as the governing concept in the construction of a theory of meaning. Putting it another way: that a declarative sentence has truth-conditions is too significant (or perhaps too trivial) a fact about it to be dispensable, for to dispense with truth-conditions would be to dispense with declarative sentences themselves.¹³

The unsatisfactory nature of the skeptical solution suggests that one of the options originally rejected by Kripke may after all be correct. This was the thesis that understanding is an irreducible, *sui generis* mental state. The failure of the dispositionalist’s account shows that there is no causal surrogate for this

state. Meaning, and normativity, are not reducible to the nonsemantic and the purely descriptive. The state of understanding is one that does not consist in being in any physical—or, for that matter, mental—state expressible in terms that do not presuppose the notion of meaning.¹⁴ Understanding is not a state that consists in some *other*, not essentially meaning-involving, kind of state. The point is amusingly brought out by Simon Blackburn's presentation of the skeptical problem. Suppose that I learnt the rule of addition yesterday, and that today I arrive at the computation $57 + 68$. If I am to be faithful to the rule I learnt, we naturally want to say, I should reach the answer 125. "I most certainly should not say that $57 + 68$ is 5. Nor of course should I say that there is more than one answer to that problem, or that the problem is indeterminate, so that there is no answer at all" (2002, pp. 33–4). But now we are told that our difficulties really begin: the skeptic "asks me to point to the fact that I am being faithful to yesterday's rule only by saying one thing [namely '125'], and not these others. And this proves hard to do" (*ibid.*, p. 34). But what is so hard about it? The skeptic has just done it! What Blackburn's skeptic meant to ask me to do, of course, was to point to *another* fact, one that could then *constitute* the first fact—the fact that I am being faithful to yesterday's rule only if I say "125." And when the challenge is put like this it seems obvious that the right response is simply to refuse to adduce any such further fact.

3 Wittgenstein on Dispositions

Probably the part of Kripke's argument that has generated most interest is his discussion of dispositions, and his claim that a dispositional account is incompetent to reconstruct meaning and understanding. I turn now to examine this point further. It is both exegetically interesting and philosophically instructive to ask first whether Kripke's argument against the dispositional account of understanding reflects Wittgenstein's own position. Kripke himself supposed that it did, but against this Colin McGinn has contended (1984, pp. 72–7) that Wittgenstein would in fact have accepted the dispositional account on the basis of considerations relating to the operation of counterfactual conditionals. Now it is a consequence of Kripke's rejection of the dispositional account of meaning that no counterfactual conditional specifying what I *would do* or *would have done* in any given unrealized situation can constitute my understanding of a rule. For given that my dispositions are finite, and that I am prone to make mistakes, what I *would do* or *would have done*, when confronted with some situation in which application of a rule is called for, may (in fact will) come apart from what I *ought to do*, or *would have been required to do*, without prejudice to my capacity really to grasp the rule. McGinn cites a passage in

which Wittgenstein seems to connect one's grasp of a rule with considerations of what one would have done in certain unrealized situations:

"But I already knew, at the time when I gave the order ['add 2'], that he should write 1002 after 1000." —Certainly; and you may even say you *meant* it then; only you should not let yourself be misled by the grammar of the words "know" and "mean." For you do not mean that you thought of the step from 1000 to 1002 at that time—and even if you did think of this step, still, you did not think of other ones. Your "I already knew at the time . . ." amounts to something like: "If I had then been asked what number he should write after 1000, I would have replied '1002.'" And that I do not doubt. This is an assumption of much the same sort as "If he had fallen into the water then, I would have jumped in after him." (2009, I, §187, translation adapted)

But this text does not show that Wittgenstein embraces a dispositional account of meaning in the sense in which that account is rejected by Kripke. The crucial point is that the examples Wittgenstein cites of hypothetical circumstances place no unusual strain on the rule-follower: the circumstances in which it is relevant to assess the counterfactual are clearly meant to be normal ones, and there is in particular no question of presenting the rule-follower with situations that challenge his finitude. So the passage tells us nothing about what Wittgenstein would have said about circumstances in which excessive strain is placed on the finitude of the understander's capacities, or which in some other way are abnormal.

It follows that, although this passage does not directly *support* Kripke's interpretation of Wittgenstein, according to which there is no general availability of a dispositional account of rule-following, we can at least say, against McGinn, that it is *compatible* with that interpretation. But the passage does suggest that Wittgenstein would have allowed that in circumstances certified as normal there may be a *close correlation* between a rule-follower's dispositions and his understanding. In fact, such a restriction to "normal" circumstances surely guarantees that there is more than a mere *correlation* between these two things: a dispositional account will be *correct* for normal circumstances. But that is a truism, not an interesting discovery: normal circumstances, we can say, are constitutively just those for which the dispositional account holds. If, in (real or hypothetical) circumstances certified as normal, I fail to deliver the right answer to the question "What is the sum of m and n . . . ?," what that shows is that I do *not*, after all, grasp the rule of addition; rather, my understanding is indeed constituted by whatever deviant-function-computing disposition my behavior manifests. Similarly, if, in (real or hypothetical) circumstances certified as normal, I fail to jump into the water to save someone, what that

shows is that (contrary perhaps to my antecedent claim) I am *not* brave. To put it another way, *normal* circumstances are those in which what anyone who grasps the rule would do coincides with what the rule says you should do, so that if you do not do it, that shows that you do not grasp the rule; it is only in *abnormal* circumstances—but this is enough to scupper the dispositionalist's story—that we permit your understanding and your actual behavior to come apart. Wittgenstein can be taken to be assenting to this truism in the quoted passage, and that without prejudice to the question whether a *general* dispositional account of rule-following is available.

In the passages immediately following the one quoted above in this section, Wittgenstein distinguishes between two senses in which meaning something by the expression of a rule may be said to determine the application of the rule in advance. One of these senses Wittgenstein rejects as illegitimate; the other he accepts. Wittgenstein does not say *how* meaning can (in the legitimate sense) determine the steps of a rule in advance: but *that* it can do so is clearly implied. What is the sense in which meaning does *not* determine the steps in advance? That is made clear in §195, and it is in the light of this passage that the difficult intervening remarks on supermechanism and the machine-as-symbol are to be understood. What emerges from §195 is that the illegitimate sense in which one might be tempted to suppose that meaning determines future applications of the rule is a *causal* one. My understanding of the rule does not drive *mechanically* (or in any other causal way) future applications of the rule. Now a dispositional account of rule-following is a causal account: if understanding a rule consisted, quite generally, in the possession of a disposition to give certain responses in relevant situations, then grasping a rule would indeed drive future applications of the rule “causally and as a matter of experience.” Wittgenstein rebuts this latter suggestion; it follows that Kripke is correct to ascribe to Wittgenstein rejection of the dispositionalist model of understanding.

Further, Wittgenstein's explicit reasons for rejecting the causal account, though they make no reference to human finitude, do draw on Kripke's other main consideration telling against that account: the need to accommodate the possibility of error. For if rule-following, and so linguistic understanding, were conceived to work causally, no theoretical room would be left for mistakes. Whatever a machine (or any causal system) does causally it does without leaving the causal order (that is a tautology); but we need to be able to make theoretical room for the possibility that a machine, while remaining an integral part of the causal order (without suddenly behaving *acausally*), may, as we say, *malfunction*. Hence we need to make room for a *noncausal* account of going wrong: a machine malfunctions just if it fails to conform to its designer's intentions; that failure is not purely a matter of its behavior as part of the causal order. And if our account of going wrong is noncausal, it follows

that our account of going right (following a rule) will also be noncausal. Our account of what a machine *does* will of course be a causal one; but our account of what a machine *ought to do* (i.e. what it is intended by its designer to do) will be noncausal, and so also nondispositional.

At this point Wittgenstein warns us against succumbing to a temptation to *sublime the machine*. We are liable to confuse the actual, nuts-and-bolts machine with what he calls the machine-as-symbol, that is, with its *blueprint*. But these are quite distinct: “the movement of the machine-as-symbol is predetermined in a different sense from the way in which the movement of any given actual machine is predetermined.”¹⁵ The movement of an *actual machine* is predetermined causally—which means, in effect, that from our current epistemic perspective there are many things that might, just as a matter of sheer logic, happen. Of course we think we know how the machine will behave. And indeed we often *do* know; our claim to know is not just bravado. But the point is this: the *basis* of our knowledge, in the cases where we have it, is inductive and experiential, not a priori or logical. There is nothing *in logic* to prevent the machine from behaving in indefinitely many ways (though no doubt not in just any way at all: this was, in essence, Kant’s point against Hume: see Strawson 1985), and so in that sense how the machine *will in fact* behave is something we have, ultimately, to wait to discover. By contrast, the movement of the *machine-as-symbol* is predetermined in just the sense in which the applications of a rule are predetermined by the rule—logically, not causally. The blueprint of a machine tells us how the machine *ought* to behave, or how it will behave *if conditions are ideal*. And, as we have seen, the appeal to ideality imports a new dimension: here, it takes us out of the causal into the logical order. For although it is possible to give a *partial* specification of ideal conditions in purely empirical terms and so without begging the question, it is only *ultimately* possible to specify them if one explicitly or by implication includes the requirement that the system in whose behavior one is interested does actually behave in the intended way in those conditions. That has somehow to be guaranteed, otherwise the conditions will not count as ideal: and the only way to *guarantee* anything in this mortal life—guarantee it so that there are *no* exceptions, and no *possibility* even of an exception—is to make it a matter of logic that the relevant conditions are met. Hence we say, truistically enough, that ideal conditions are those in which an actual machine, built following the blueprint (the machine-as-symbol), behaves just as the blueprint predicts.

So far so good, but danger lurks at this point. Wittgenstein thinks that we are liable to confuse the actual machine with the machine-as-symbol, and so to come up with the metaphysical monstrosity of a *supermachine*, a machine that is *both* actual *and* ideal. The idea of a supermachine is the idea of an entity that is both an ordinary, physical object—something that we could in principle build with bits of metal and plastic—and also a rule-follower that cannot

malfunction, that is guaranteed to deliver the right results for any input, in any circumstances. It is an *empirical* machine that is *logically* guaranteed to follow a rule. The reason why the idea is a monstrosity is—at bottom—that the logical cannot be reduced to the causal. The application of a *rule* (i.e. how the rule should be applied, or, as we often more simply say, how it applies) cannot misfire because, as a matter of logic, a misfiring application of a rule would be a *misapplication* of it—that is, it would *not* be an application of it at all. But any empirically real *mechanism* can misfire: that is, any such mechanism can perform in a way in which it was not intended to perform. The concept of a supermachine arises as a result of what Wittgenstein calls “the crossing of different pictures” (2009, I, §191)—namely, the logical and the causal.

In the *Philosophical Investigations*, Wittgenstein’s interlocutor finds it odd that there could be a noncausal way of anticipating the future.¹⁶ That picks up a remark made in *Zettel*:

How strange: it looks as though, while a physical (mechanical) form of guidance could misfire and let in something unforeseen, a rule could not do so! As if a rule were, so to speak, the only reliable form of guidance. But what does guidance’s not allowing a movement, and a rule’s not allowing it, consist in?—How does one know the one and how the other? (1989c, §296, translation adapted)

The strangeness here derives from our thinking of the rule as a form of *guidance*. For in fact the rule does not *guide* its applications, it *characterizes* them. If we thought of it as guiding its applications, it would have to be a super-strong form of guidance that simply could not misfire (unlike an ordinary mechanism, which can). But that takes us to the incoherent idea of a supermachine, our reflections on which have told us that the relation between a rule and its applications is logical, not causal.¹⁷ In the light of this exegetical discussion, let us now review the dispositional account of understanding, and see whether it can, as a number of commentators have thought, be defended or in some way modified so as to withstand Kripke’s objections to it.

4 Dispositions and Normativity

According to the dispositionalist’s account, understanding a word consists in having a set of appropriate dispositions to use the word in certain ways in particular circumstances. The basic problem with this account, according to Kripke, is that it fails to close the gap between the descriptive and the normative. If I have a disposition to apply a word in certain ways in particular

contexts, we can use a *description* of that disposition to predict what I *will in fact* do; but this characterization falls crucially short of having implications for what I *ought* to do. Kripke puts the point as follows:

a dispositional account misconceives the sceptic's problem—to find a past fact that *justifies* my present response. As a candidate for a “fact” that determines what I mean, it fails to satisfy the basic condition on such a candidate, . . . that it should *tell* me what I ought to do in each new instance. (1982, p. 24)

This passage gives rise to a number of qualms, mainly attaching to the phrase “tell me what I ought to do.” First let us ask: in what sense *ought* I to do what the rule tells me? If I am contemplating a problem of the form “What is $m + n$?” in what sense *ought* I to respond with the *sum* of m and n , as opposed to giving the output of some other function, or of no function at all? Can we not imagine circumstances in which the right thing for me to do would be to give an incorrect answer, or no answer at all? If there is some sort of obligation on me, perhaps of a *prima facie* nature, to respond to the question with the sum of m and n , is this a semantical obligation, a moral one, a prudential one, or what? Does the idea of a semantical obligation make sense? These and similar questions have generated an extensive secondary literature.¹⁸ But without broaching the question whether there is some kind of categorical obligation on me to use words in a particular way—for example, whether I ought, in some *prima facie* or even unconditional sense, to tell the truth—for present purposes we can focus on a smaller point, indeed one that has passed for a platitude in the literature.

This is the point that the rule—the meaning of a word—has normativity built into it in the sense that it divides possible applications into *correct* and *incorrect* uses.¹⁹ As Paul Boghossian puts it:

Suppose the expression “green” means *green*. It follows immediately that the expression “green” applies *correctly* only to *these* things (the green ones) and not to *those* (the non-greens). The fact that the expression means something implies, that is, a whole set of *normative* truths about my behaviour with that expression: namely, that my use of it is correct in application to certain objects and not in application to others.²⁰

Those who have defended (or attacked) the doctrine of the normativity of meaning have generally associated something stronger than this platitude with that doctrine, but the platitude will suffice perfectly well for our (and Kripke's) purposes. For it is exactly the normativity of meaning in this modest sense that dispositionalism cannot accommodate.²¹ This was the moral of

the previous section: the rule governs its applications logically, not causally; so a dispositionalist story cannot reconstruct from its—necessarily merely causal—materials the real nature of the rule. In connection with the point that, for our purposes, what counts as the correct use of the word “green” is just its application to objects that are green, the following should be noted. We utter such sentences as “You have used the word ‘vermilion’ incorrectly—to mean something that it does not mean.” This kind of locution is in good order provided it is glossed along the lines of “Your understanding of the meaning of the word ‘vermilion’ does not agree with its actual meaning.” What a speaker cannot do, of course, is *change the meaning of the word*; that is, you cannot use the word “vermilion” in such a way as to make it mean, either on your lips or in general, something other than its actual meaning—*blue*, say, or *square*. No individual has that power over the meanings of words: “vermilion” means what it means in the public language irrespective of your or my individual uses of it, and there is nothing you or I can do to change that.²² The most that individual speakers can achieve, so to speak, by way of making a mistake with that word, is (apart from using it ungrammatically) to apply it to objects *that are not vermilion*.

This is a matter that occasionally confuses commentators. For instance, Claudine Verheggen remarks that “my meaning rain by ‘rain’ always gives me a reason to use (commits me to using) the term in certain ways, . . . I do not have the option to be indifferent to the fact that I mean rain by ‘rain’” (2011, p. 563). But this formulation is misleading inasmuch as it suggests that I have a choice—that I have a room for maneuver in the way I deploy the word “rain”—which in reality I do not have. To say that I have a reason, or am committed, to use the word “rain” to mean *rain* suggests that I could choose *not* to use that word in accordance with its meaning; but of course I have no such freedom. At least, I do not have that freedom so long as I speak English. I can of course set up a personal code in which “rain” means anything I like, but so long as I participate in the public language, I do not have the option of using “rain” otherwise than in accordance with its actual meaning. The most that the general reasons and commitments hereabouts, of which Verheggen talks, could possibly amount to—given that there is no general obligation *not* to devise personal codes—would be an obligation to tell the truth. Whether I have that obligation, and what its basis is if I do, is, as I have said, a question that I am setting aside, since for our purposes it is quite sufficient that there is a distinction between correct and incorrect applications of a rule, whether or not anything obliges me to apply rules correctly. That distinction already refutes dispositionalism, given that if one tried to reconstruct the idea of the correct application of a rule causally, one would be forced into positing the metaphysical monstrosity of a supermachine. Dispositionalism cannot reconstruct *the rule*, and insofar as I, an individual language user, latch on to the

rule in understanding, it cannot reconstruct my grasp of the rule either. I shall return to this point in the next section.

There is a further qualm attaching to the phrase “tell me what I ought to do” in the passage from Kripke quoted at the beginning of this section, this time one that focuses not on the “ought” of that phrase but on the word “tell.” One might complain that the locution “tells me what I ought to do” reintroduces precisely the discredited notion of rule-following as having a recipe, in the form of some mentally accessible linguistic or pictorial image, which guides my use.²³ And Wittgenstein, as we have seen, indeed rejects that idea of what rule-following consists in again and again: “One does not feel that one has always got to wait upon the nod (the whisper) of the rule. On the contrary, we are not on tenterhooks about what it will tell us next, but it always tells us the same, and we do what it tells us”²⁴—which is as much as to say that it does not *tell* us anything, for if following a rule were like listening to the whisper of an inner (or outer) voice, we *would* be on tenterhooks about what it was going to say next. It is true that Kripke’s way of expressing the objection to dispositionalism in the cited passage is incautious in the way the above complaint identifies. But Kripke’s particular way of articulating his objection should not be pressed too hard, because the fundamental difficulty survives reformulation in a more hygienic locution.²⁵ As I put it above, the essential objection to dispositionalism is this: having a disposition falls crucially short—not of *telling me*, but—of *having implications* for what I ought to do. Possessing a disposition does not even have *any* implications for what I ought to do, let alone the *right* implications.

So the basic objection to dispositionalism is that it cannot recapitulate the state of understanding, which is a normative state, that is, a state that has implications for how I should behave, not (or not simply) for how I will behave. It may be that I will in fact diverge from the rule—this could be for any number of reasons, such as inattention, exhaustion, because the situation is too complicated for me to grasp, and so on—but my failure to apply the rule correctly need *not* mean that I have failed to grasp that rule, but some deviant alternative instead, as the dispositional account would imply. It may simply be that the circumstances are unpropitious, and hence that I make a mistake, as we all are prone to do, in spite of the fact that I do grasp the rule. So one reason why the dispositionalist’s account fails is that it can make no room for the notion of a *mistaken* application of the rule. Whatever I do will constitute an activation of the relevant disposition, but if we want to say, as we surely do, that on some occasions I apply the rule wrongly, *even though I do grasp the relevant rule*, we cannot at the same time affirm that to grasp the rule is to possess just that disposition. Consider again our example. I know how to add: that is, I grasp the rule associated with the sign “+.” When I add, I activate a disposition: call it *D*. But does my understanding *consist* in my

having that disposition? No. If, when I am exhausted, I am asked to perform an extremely large and complicated addition, I will very likely make a mistake. In so doing I activate *D*, and if someone knew enough about my physical state and the attendant circumstances, he or she would be able to predict my mistake. But does my making this mistake show that my understanding was in fact not of the rule associated with “+” but of another, deviant rule—the one that activating *D* actually caused me to perform? No, again. My *understanding* of “+” indeed targets addition, and not the nonstandard operation I actually perform in these exceptional circumstances. But my *disposition* is to perform the deviant operation (*D* turns out to be a disposition to add in normal circumstances, but to perform deviant operations in certain abnormal ones). Hence my understanding does not consist in having that disposition.

5 Can Dispositionalism Be Rescued?

One might try to save the dispositional account by making either or both of two moves.²⁶ First, one might adduce dispositions to recognize mistakes, and argue that, when these dispositions are factored in to our theoretical account of understanding, grasping a rule will indeed consist in having dispositions (of the right sort).²⁷ But, though this move embellishes the saga with an extra twist, it does not neutralize the basic objection to dispositionalism. These more complicated dispositions too can, under readily imaginable circumstances, deliver up wrong results: that is, there will be situations in which the right thing to say will be that the subject indeed understands the relevant word, but fails to apply the rule correctly, and also fails to activate the relevant set of mistake-detecting dispositions. Equally, there will be situations in which the mistake-detecting dispositions are activated, but wrongly: there was in fact no mistake to be corrected. It is no use trying to narrow down the range of situations that are to count as genuinely activating the relevant dispositions by inserting a *ceteris paribus* clause, or similar—this is the second way in which one might try to shore up the dispositional account—thereby seeking to exclude pretenders. The point of this maneuver would be to ensure that genuine activations of the disposition exactly coincided with the correct applications of the rule. But the only way to guarantee the requisite coincidence would be to *appeal* to the correct conditions as a means of narrowing down the available dispositions to just the acceptable ones. For nothing purely *internal* to the dispositions themselves could serve to narrow the range. But then it would no longer be the case that dispositions *as such* were constituting understanding: dispositions as such would not be delivering an independent criterion of understanding; rather, states independently certified as states of understanding would have to be adduced in order to isolate the relevant dispositional states.²⁸

That we should avoid giving a theoretical account of understanding that imports two different and possibly conflicting criteria for its obtaining is a point stressed by Wittgenstein:

If one says that knowing the ABC is a state of mind, one is thinking of a state of an apparatus of the mind (perhaps a state of the brain) by means of which we explain the *manifestations* of that knowledge. Such a state is called a disposition. But it is not unobjectionable to speak of a state of the mind here, inasmuch as there would then have to be two different criteria for this: finding out the structure of the apparatus, as distinct from its effects. (2009, I, §149)

Wittgenstein here rejects a physicalist account of understanding in general, on the basis that we would then have two different criteria of understanding: knowledge of the physical construction of the apparatus (the brain), and knowledge of what the understander says and does. Although Wittgenstein does not spell out the point, he presumably has in mind the thought that this constitutes a *reductio ad absurdum* of the physicalist view for the following reason: if there were two different criteria of understanding—as, assuming physicalism, there would have to be, since the ordinary behavioral criteria of understanding would not simply lapse—then it would in principle be possible for application of these criteria to deliver divergent results on the question whether someone understood a word (otherwise they would not really be *distinct* criteria). We might, for example, decide on some occasion that the presumed physical basis for understanding was in place in a given subject, but then examination of that subject's behavior might reveal that he in point of fact did not understand the relevant word. Of course, faced with such a situation, rather than abandoning physicalism we might choose to expand our physical definition of the understanding in question. But clearly even the expanded definition would be liable to the same failing. And no matter how far we expanded it, the same theoretical possibility would remain. That is, it would always be in principle possible for someone to meet the physical criteria for understanding while failing, in behavior, to display understanding, and in such a situation we would always allow behavioral considerations to trump physical ones. It would not help here to conceive of the physical criteria as having the form of an open-ended disjunction of alternative physical states, any of which might underlie the behavioral condition of understanding. The crucial phrase in the last sentence was “open-ended.” The disjunction, in order to be adequate, would indeed *have* to be open-ended, but that would mean that the physical specification would not be serving as an *independent* criterion of understanding. Admission to the disjunction of physical states would be *dependent* on what went on at the

behavioral level: explanatorily speaking, specification of behavior would be in the driving-seat, and specification of underlying physical conditions would tag along behind. So there could be no *reduction* of understanding to a set of physical conditions.

Recall that Kripke faulted the dispositional account not only on the basis that it could not accommodate mistakes but also because it failed to accommodate human finitude. Boghossian objects to the point about finitude on the grounds that we constantly make assessments about what would happen in certain never-to-be-realized situations, and in doing so we permit ourselves to operate with *idealizations* of those situations: we permit ourselves to operate with a restricted range of parameters (2002, pp. 164–6). Thus we make predictions about how gases *would* behave under conditions that we will never in fact encounter. We allow ourselves to treat relevant counterfactual conditionals as true, even though we have, in one sense, no idea how gases would *really* behave were those ideal conditions to obtain: it is at least epistemically possible that, if those conditions were *actually* realized, gases would behave in quite unforeseen ways. But in speaking of how gases *would* behave in ideal conditions, we allow ourselves to exclude possible sources of interference, and we take ourselves to know what the effect of excluding those interferences would be. So, in assessing the relevant counterfactual conditionals, we achieve the results we want by setting the parameters in the way we want. Why, Boghossian asks, should the same not be true of conditionals specifying how we *would* apply words in situations with which, because of our finite nature, we will never in fact be confronted? That is, why should we not so set the parameters of ideal conditions that our specification of how we would behave in such ideal conditions exactly matches our understanding? The answer lies to hand, and indeed is expressed by Boghossian himself. Of the idealizations in question he remarks: “Obviously, only certain idealizations are permissible; and also obviously, we do not now know which idealizations those are. The set of permissible counterfactuals is constrained by criteria of which we currently lack a systematic account” (ibid., p. 166). But it is no accident that we do not have this systematic account. The reason why we lack it is that the only way to construct it would be to help ourselves illegitimately to the concept of meaning itself. That is, any idealizations we could form of how we *would* behave if we had infinitary dispositions would have to be so gerrymandered that they exactly mimicked the correct application of the rule. The idealizations could not be specified in a nonquestion-begging way.²⁹

Suppose I contemplate a massive computation problem, so large that I would not only die long before I had completed it but also go mad owing to the excessive demands placed on my mathematically modestly endowed brain. And suppose now that *under ideal conditions*—that is, in this case,

given a suitable expansion of my brain-capacity, a sufficient lease of life, and assuming no failure of concentration, absence of interference, and so on—I would carry out the computation correctly. What constitutes the *ideality* of the conditions? It is no doubt possible to give a *partial* specification of these without deploying, or presupposing appeal to, the concepts of meaning or understanding in a question-begging way, and I have just done so. But that cannot be the whole story. Suppose we realize all such conditions that we can think of, and suppose I nevertheless fail to perform the computation correctly. One possible explanation is of course that I have not actually grasped the rule (so there would be no necessary mismatch between dispositions and understanding here). But another perfectly genuine possibility is that the conditions are not yet ideal, and, if we suppose that I *have* grasped the rule, the nonideality of the conditions can consist in nothing else than that I *do not* under those conditions compute correctly. My failure to do so precisely shows that the conditions were after all not ideal. Recourse to the notion of ideal conditions smuggles in an illicit appeal to the notion of the correct application of the rule.

The give-away (which the reader may have spotted) in my specification above of ideal conditions for performing a massive computation—to remind you, it went: “given a suitable expansion of my brain-capacity, a sufficient lease of life, and assuming no failure of concentration, absence of interference, and so on”—was of course that seemingly innocuous phrase “and so on.” This phrase is, in this context, very far from innocent: it holds place for an *open-ended disjunction*. But that means, as we have seen, that the specification of ideal conditions is not genuinely *independent*. For the items that gain admission to the disjunction and so count as components of our specification of ideal conditions can only *ultimately* be fixed by stealing a sidelong glance at actual practice under various conditions—by eking out the descriptive with the normative, by allowing our characterization of what we *would* do in endless possible situations to be fixed by an understanding of what we *ought to* do in those circumstances. The list is supposed to reconstruct understanding, but since no finite list can do this, the trick of appealing to an infinite list that does reconstruct it can only be turned by drawing on an anterior conception of understanding in fixing the relevant conditions. But then the list of descriptive conditions has no independent authority for the purposes of fixing the nature of understanding. The issue that we have highlighted here is one that quite generally foils attempts to naturalize semantics.³⁰

Of course, my understanding must in some way be embodied in a brain state, and brain states *are* dispositional states; the brain works in a causal way. But, as we have said, the *rule* does not work causally but logically; and there is an absolute gap between the logical and the causal. So it follows that, insofar as an understander genuinely grasps a rule, his or her state of understanding,

notwithstanding its presumed embodiment in a brain state that is a dispositional state, is not *itself* a dispositional state, or in general any state of which a causal account of its operation would be correct. The rule does not drive its applications causally, as we have seen, because to do that it would need to be a supermachine; and that idea is fantastic. There cannot be an empirical object that is guaranteed as a matter of logic to behave according to a rule. And of course the idea of a supermachine is merely a philosopher's fantasy. In practice, as Wittgenstein says (1977, I, §193), we do not forget the possibility that any given machine may malfunction; we do not turn our judgment over to a particular computer that we have programmed to add and say that addition is whatever it outputs, come what may. So Warren Goldfarb's reductionist fantasy of a "future physiological psychology" that revealed states of competence in the brain that were such that, if untrammelled by interfering factors, they would "always cause correct responses" (2002, p. 97) can be ruled out on a priori grounds: such states would be supermachines, and in practice we would never put our arithmetical judgment (say) in thrall to the outputs of a given mechanism, overlooking or discounting the possibility of malfunction. In an effort to make the fantasy more realistic, Goldfarb supposes that "spoiler" brain states are also discovered and that these account for errors in rule-following. But you do not make the hypothesis of supermachines less absurd by hooking them up to interference.

Now human rule-followers are fallible. The understanding of any given rule-follower may, in respect of any particular rule, be deviant. Perhaps those people walking past in the street outside—perhaps you and I—do *not* grasp the rule of addition, but rather a deviant function that happens to coincide with addition over relatively small and relatively easy inputs. However—and here the discussion necessarily takes a transcendentalist turn—that cannot be the *general* situation. People have thought about the rule of addition since time immemorial. It cannot be the case that what they have really been talking and thinking about is not actually addition, but rather a deviant, star-like function (or a set of different such functions). We, as a mathematically trained community, take ourselves to be able to talk and think about *addition*—that particular function, with *all* its applications. We know what we are talking about; so our understanding has no merely causal reduct. And, insofar as we permit any individual thinker to latch on to the same function and grasp the same rule, it follows that that thinker's understanding likewise has no dispositional, or in general causal, reduct. So the key move in the refutation of dispositionalism is the point that *the rule* has no causal surrogate; given that fact, it follows that, insofar as we allow people to grasp the rule, their understanding cannot consist in a merely causal state, such as a disposition, either. In grasping a rule we—individually and collectively—are put in touch with the ideal, in a way that goes beyond the merely causal.

Here it might be objected—to develop a point that was raised earlier—that the ascription of ordinary causal dispositions to things, for example, solubility to salt or fragility to glass, also involves the ideal: for, in predicating solubility of salt, we are not saying that salt will dissolve (in water) come what may, but only that it will dissolve *ceteris paribus*, and that proviso involves an implicit idealization.³¹ But the point is this. We can indeed define, say, the *solubility of salt in water* in terms of certain idealizations, in such a way that it is logically guaranteed that, in ideal circumstances, salt really does dissolve in water. We can do this without triviality, *if* salt dissolves in water in a range of standard and independently identifiable situations: the proviso is crucial, for otherwise there would be nothing to stop *anything* from counting as soluble (fragile, etc.), so long as the idealizations were suitably gerrymandered. But we cannot go through this definitional procedure for the *concept of dissolving* itself. *What it is for something to dissolve* cannot be identified with the behavior of any physical system without either trivialization (if that system is identified in a question-begging way) or falsity (if it is not). Similarly, *what the sum is* of any given two input numbers cannot be identified, without trivialization or falsity, with the outputs of any particular empirical process or mechanism. We might think of the difference like this: something counts as soluble in water if it has one of a range of dispositions—an infinite set, in fact, containing dispositions that coincide in standard situations (in these the substance has to dissolve in water to count as water-soluble) and diverge in nonstandard ones (in these the substance does not have to dissolve in water without its ceasing to count as water-soluble); by contrast, the rule of addition is not a set of functions that coincide over only part of the domain of pairs of numbers and possibly diverge elsewhere, but is a unique function over the whole finite and transfinite domain. No causal surrogate for the rule could deliver that result. And, so far as we manage to grasp the rule, the same goes for us rule-followers.

6 Conclusion

In conclusion, let me clarify a point about the way dispositions have figured in the above discussion that may have been worrying the reader. For during this discussion two different conceptions of what dispositions are have in fact emerged. Indeed, the antinomy between these two conceptions pervades the literature on rule-following and normativity, and it is time to bring it into the open. We can present the dispositionalist with a dilemma, induced by the question: it is possible to act or behave in a way that does *not* accord with one's dispositions? On one conception of the nature of dispositions, this is

not possible: whatever one ends up doing, *that* is what one was disposed to do.³² This is the conception of dispositions that Kripke presupposes and that I was presupposing in the argument of my last paragraph in Section 4. If dispositionalists adopt this conception, then I take the argument I gave in that paragraph to be decisive against dispositionalism. On the other hand, talk of *ceteris paribus* clauses and idealizations imports a conception of dispositions according to which one may act *against* one's dispositions: on this conception, one may be disposed to add even if one makes mistakes.³³ Should the friends of dispositionalism opt for this conception of what dispositions are, then I regard the final paragraph of Section 5 as refuting their view. Either way, I take it that dispositionalism fails.³⁴

Notes

- 1 Compare Miller, Introduction to Miller and Wright 2002, p. 2; Wright 2002, p. 109.
- 2 *Contra* Wright 2002, pp. 119–20.
- 3 1977, I, §§198–202, 433, 503–4; compare 1969, p. 34.
- 4 I think it makes better sense to structure Kripke's argument in this hierarchical way, with the normativity objection functioning as a genus of which the mistake and finiteness objections are species, than to present the three objections as on a level (so Ginsborg 2011a, p. 228).
- 5 Miller, Introduction to Miller and Wright 2002, p. 3 n. 6; Wright 2001, p. 96.
- 6 Compare Wright 2001, pp. 104–6; Boghossian 2002, pp. 160–1. Wright argues that global nonfactualism is unstable; Boghossian is not so sure. I agree with Wright, though not for the reasons he gives, but for the ones I give in the text.
- 7 But I suggest that it is an innocent, constituting regress: see my 2008, ch. 6.
- 8 On Wittgenstein's approach to this issue, see McDowell 1998, Essay 11 at p. 234 and *passim*; Verheggen 2003, pp. 305–7.
- 9 Compare Wittgenstein 1977, I, §§240–2.
- 10 1998, p. 252; compare Goldfarb 2002, pp. 103–4.
- 11 Wittgenstein 1989a, §IV–28.
- 12 McDowell, 1998, p. 252. Although I am in agreement with McDowell's basic stance in this passage—his opposition to an anti-realism about meaning—it should be noted that his argumentation on pp. 252–3 (with n. 49) commits the fallacy of composition, something that McDowell is elsewhere prone to do: see my 2006, pp. 124–5; 2013, §3.
- 13 Compare Boghossian 2002, pp. 161–2.
- 14 See here Haddock 2012, pp. 153–4, 160–1, relying on McDowell 1998, esp. essay 12.
- 15 2009, I, §193 (translation adapted).
- 16 1977, I, §195.
- 17 It is interesting to read these sections of *Philosophical Investigations* alongside those parts of the *Remarks on the Foundations of Mathematics* where Wittgenstein distinguishes between calculation and experiment, and where a similar point is being made.
- 18 See, for example, Glüer 1999; Wikforss 2001; Boghossian 2003, 2005a, b; Hattiangadi 2006, 2007; Whiting 2007, 2009; Glüer and Wikforss 2009; Verheggen 2011.
- 19 See, for example, McDowell 1998, pp. 221–2 with n. 4; Glüer 1999, pp. 38, 118–20; Blackburn 2002, pp. 28–9.

- 20 2002, p. 148. Boghossian adds: "This is not, as McGinn would have it, a relation between meaning something by an expression at one time and meaning something by it at some later time; it is, rather, a relation between meaning something by it at some time and its *use at that time*" (ibid.). But that is not right either: in fact, it is a relation between meaning something by it at some time *t* and its use at any time at or after *t*.
- 21 Compare Verheggen 2011, pp. 564–5; Ginsborg 2011a, pp. 241–4.
- 22 See here my 2008, §85.
- 23 See here Wikforss 2001, p. 217; Ginsborg 2011b, p. 157; 2012, pp. 128–9, 138–9, 144.
- 24 1977, I, §223 (translated Anscombe 1958).
- 25 Compare Haddock 2012, pp. 151–2, 155–6.
- 26 Compare Kripke 1982, pp. 30–2.
- 27 Compare Blackburn 2002, pp. 35–6. Similar remarks as I make here about dispositions to identify mistakes would apply against Ginsborg's strategy of appealing to dispositions to regard one's behavior as appropriate or inappropriate: 2011a, pp. 244–5; 2012, pp. 137–8.
- 28 Compare Boghossian 2002, pp. 176–7, 186.
- 29 Compare Boghossian 2002, p. 186.
- 30 Compare Loewer 1997, esp. p. 114.
- 31 Compare Wikforss 2001, p. 208; Blackburn 2002, p. 35; Forbes 2002, p. 23; Ginsborg 2011b, pp. 158–60.
- 32 So, for example, Brandom 1994, p. 29; compare Glüer 1999, p. 100.
- 33 So, for example, Ginsborg 2011a, p. 245.
- 34 Many thanks to Barry Dainton for his helpful comments on an earlier version of this paper.

Bibliography

- Blackburn, S., 2002. "The Individual Strikes Back." In A. Miller and C. Wright, eds, *Rule-Following and Meaning*. Chesham: Acumen, pp. 28–44.
- Boghossian, P., 2005b. "Is Meaning Normative?" In C. Nimtz and A. Beckermann, eds, *Philosophie und/als Wissenschaft*. Paderborn: Mentis, pp. 205–18.
- , 2005a. "Rules, Meaning, and Intention—Discussion." *Philosophical Studies*, 124, pp. 185–97.
- , 2003. "The Normativity of Content." *Philosophical Issues*, 13, pp. 31–45.
- , 2002. "The Rule-Following Considerations." In A. Miller and C. Wright, eds, *Rule-Following and Meaning*. Chesham: Acumen, pp. 141–87.
- Brandom, R., 1994. *Making it Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, MA: Harvard University Press.
- Forbes, G., 2002. "Scepticism and Semantic Knowledge." In A. Miller and C. Wright, eds, *Rule-Following and Meaning*. Chesham: Acumen, pp. 16–27.
- Gaskin, R., 2013. "When logical atomism met the *Theaetetus*: Ryle on naming and saying." In M. Beaney, ed., *The Oxford Handbook of the History of Analytic Philosophy*. Oxford: OUP, ch. 29.
- , 2008. *The Unity of the Proposition*. Oxford: OUP.
- , 2006. *Experience and the World's Own Language: A Critique of John McDowell's Empiricism*. Oxford: Clarendon.

- Ginsborg, H., 2012. "Meaning, Understanding, and Normativity." *Proceedings of the Aristotelian Society*, supplementary volume 86, pp. 127–46.
- , 2011b. "Inside and Outside Language: Stroud's Nonreductionism about Meaning." In J. Bridges et al., eds, *The Possibility of Philosophical Understanding: Reflections on the Thought of Barry Stroud*. Oxford: OUP, pp. 147–81.
- , 2011a. "Primitive Normativity and Skepticism about Rules." *Journal of Philosophy*, 108, pp. 227–54.
- Glüer, K., 1999. *Sprache und Regeln: Zur Normativität von Bedeutung*. Berlin: Akademie Verlag.
- Glüer, K., and Wikforss, A., 2009. "Against Content Normativity." *Mind*, 118, pp. 31–70.
- Goldfarb, W., 2002. "Kripke on Wittgenstein on Rules." In A. Miller and C. Wright, eds, *Rule-Following and Meaning*. Chesham: Acumen, pp. 92–107.
- Haddock, A., 2012. "Meaning, Justification, and 'Primitive Normativity.'" *Proceedings of the Aristotelian Society*, supplementary volume 86, pp. 147–74.
- Hattiangadi, A., 2007. *Oughts and Thoughts: Rule-Following and the Normativity of Content*. Oxford: Clarendon.
- , 2006. "Is Meaning Normative?" *Mind and Language*, 21, pp. 220–40.
- Kripke, S., 1982. *Wittgenstein on Rules and Private Language*. Oxford: Blackwell.
- Loewer, B., 1997. "A Guide to Naturalizing Semantics." In B. Hale and C. Wright, eds, *A Companion to the Philosophy of Language*. Oxford: Blackwell, pp. 108–26.
- McDowell, J., 1998. *Mind, Value, and Reality*. Cambridge, MA: Harvard University Press.
- McGinn, C., 1984. *Wittgenstein on Meaning*. Oxford: Blackwell.
- Miller, A. and Wright, C., eds, 2002. *Rule-Following and Meaning*. Chesham: Acumen.
- Strawson, P. F., 1985. "Causation and Explanation." In B. Vermazen and M. Hintikka, eds, *Essays on Davidson: Actions and Events*. Oxford: Clarendon, pp. 115–35.
- Verheggen, C., 2011. "Semantic Normativity and Naturalism." *Logique et Analyse*, 216, pp. 553–67.
- , 2003. "Wittgenstein's Rule-following Paradox and the Objectivity of Meaning." *Philosophical Investigations*, 26, pp. 285–310.
- Whiting, D., 2009. "Is Meaning Fraught with Ought?" *Pacific Philosophical Quarterly*, 90, pp. 535–55.
- , 2007. "The Normativity of Meaning Defended." *Analysis*, 67, pp. 133–40.
- Wikforss, A., 2001. "Semantic Normativity." *Philosophical Studies*, 102, pp. 203–26.
- Wittgenstein, L., 2009. *Philosophical Investigations*. German text with English translation by G. E. M. Anscombe, P. M. S. Hacker, and Joachim Schulte. Revised 4th edn. Oxford: Wiley Blackwell.
- , 1989c. *Zettel*, ed. G. E. M. Anscombe and G. H. von Wright, in *Über Gewissheit* Frankfurt am Main: Suhrkamp, pp. 259–443. Translated by G. E. M. Anscombe. 2nd edn. Oxford: Blackwell, 1981.
- , 1989b. *Über Gewissheit*. Edited by G. E. M. Anscombe and G. H. von Wright. Frankfurt am Main: Suhrkamp, pp. 113–257.
- , 1989a. *Bemerkungen über die Grundlagen der Mathematik*. Frankfurt am Main: Suhrkamp. Translated as *Wittgenstein on the Foundations of Mathematics*. 3rd edn. Edited by G. H. von Wright, R. Rhees, and G. E. M. Anscombe. Oxford: Blackwell, 1978.

- , 1977. *Philosophische Untersuchungen*. Edited by G. E. M. Anscombe, G. H. von Wright and R. Rhees. Frankfurt am Main: Suhrkamp. Translated as *Philosophical Investigations*, G. E. M. Anscombe. Oxford: Blackwell, 1958.
- , 1969. *The Blue and Brown Books*. 2nd edn. Oxford: Blackwell.
- Wright, C., 2002. "Critical Notice of Colin McGinn's *Wittgenstein on Meaning*." In A. Miller and C. Wright, eds, *Rule-Following and Meaning*. Chesham: Acumen, pp. 108–28.
- , 2001. *Rails to Infinity: Essays on Themes from Wittgenstein's Philosophical Investigations*. Cambridge, MA: Harvard University Press.
- , 1980. *Wittgenstein on the Foundations of Mathematics*. London: Duckworth.

19 Philosophy of Science

James Ladyman

Philosophy of science was said by Quine to be philosophy enough. One does not have to go that far to concede that large parts of the subject matter of other parts of philosophy overlap with or are encompassed by it. For example, questions in the philosophy of psychology concern central philosophical problems about the mind-body problem, consciousness, and free will; the philosophy of economics treats philosophical issues about rationality and practical reason; the metaphysics of space, time, and matter is addressed by philosophers of physics; and even aesthetics has some overlap, not just through work on aesthetic virtues of theories, but because of a common interest in the nature of representation. Epistemology is approached in philosophy of science via formal methods such as Bayesianism, classical statistical inference, and epistemic logics, as well as being the subject of theories of the scientific method. Philosophy of science includes ethical and political reflection applied to or mediated by consideration of scientific theories; questions arise about the virtue of scientists, the social organization of science, and the link between scientific ethics and epistemic success. Hence, it is very difficult to give a comprehensive guide to the philosophy of science that is not also a guide to the rest of philosophy. The corresponding problem is that philosophy of science is also nearly as broad as science itself. Most philosophers of science learn something of one or more sciences, and many specialize sufficiently to become veritable experts. Their work may be co-written with scientists and some of it appears in science journals. The enormous growth in the range and depth of scientific disciplines means there are now philosophers of science working in every field from cosmology to ecology, and from archaeology to psychiatry. There is also a very close link between philosophy of science and logic and philosophy of mathematics. Philosophers of science are usually trained in formal logic and have a reasonable grasp of mathematical logic. Many have a deep knowledge of logic, probability theory, and other parts of mathematics, and increasingly of computer science and computational methods.

The subject came to maturity at the turn of the twentieth century with the work of Pierre Duhem and Henri Poincaré, the birth of mathematical logic, and the subsequent advent of logical empiricism, and the school of Karl Popper.

However, much earlier work remains of relevance, especially that of Bacon, Descartes, Hume, Kant, Whewell, J. S. Mill, Comte, Mach, and Peirce. The writings of great scientists, especially Newton, Darwin, and Einstein, have also profoundly influenced the subject. The core issues that defined the work of the logical empiricists, as well as Popper and his followers, remain prominent. There remains the problem of understanding the structure of scientific theories and the nature of the scientific method, and the application of formal logic to solve it. The relationship between theory and evidence, and induction, were intensively studied by Carnap and Reichenbach. Science itself has been utterly transformed since the early twentieth century by computation and mathematical statistics. The study of scientific methodology must now engage both with a huge body of work in Bayesian statistics, and the extensive use of models and simulations in science. Correspondingly, the familiar questions from the history of philosophy about scientific realism, natural necessity, causation and empiricism versus rationalism find new expression in contemporary debates, and the technicalities of mathematical biology or computational linguistics are studied to address them.

It is common to treat the philosophy of social science separately but I have not done so here. There is a lot of philosophical discussion in the history of science literature but I have not included references to it. In what follows I will somewhat arbitrarily divide philosophy of science into scientific methodology, metaphysics of science, epistemology of science, and philosophy of the sciences.¹

1 Scientific Methodology

The word “science” derives from the Latin “scientia” meaning knowledge. It is easily neglected that when the ancients and early moderns pondered the nature of knowledge and how to arrive at it they had very much less of it than we do, and there was no reason to think that theoretical science could put people on the Moon and tell us our position on the globe to within meters with a handheld device. It is easy to take for granted the idea of a systematic method of knowledge and the technology based on it giving us extraordinary power to manipulate the world. Furthermore there are many features of contemporary science that could not have been foreseen and which give it its extraordinary accomplishments. Because of its power and its proven reliability in establishing facts and changing them, we write reference to science into the law, medicine, and public policy as well as directly and/or indirectly into individual practical reasoning. Debate about the rationality of science and the demarcation between science and pseudo-science and nonscience is of great social and political importance. Climate change, immunization, and

complimentary and alternative medicine are all subjects that attract intense media and public interest and controversy, and all raise fundamental issues in the philosophy of science about theory choice and testing, and scientific rationality. Scientific status is hotly and rightly contested and the current debates about the scientific credentials of homeopathy recall earlier discussions about astrology, Marxism, and psychoanalysis that defined the classic discussions in twentieth-century philosophy of science of “the demarcation problem”: what is science and how do we tell genuine science from nonscience, bad science, and pseudo-science?²

These questions are most commonly answered with an account of the scientific method. Since its inception modern science has been regarded as constituted not by its product, but by its techniques and procedures, and its rules of inference and testing. Bacon argued for the “experimental method” at the heart of the orthodox empiricist epistemology of science, referring not only to gathering of data about naturally occurring phenomena but also to the systematic construction of artificial situations from which to infer the workings of nature. His works on philosophy of science also presciently argued that science should be collaborative and socially organized and inspired the founders of the Royal Society. His proposed new logic for the acquisition of knowledge had induction, in the form of generalization from many observations at its core. However, he has also been credited with the idea of a crucial experiment to choose between theories and the idea of hypothetico-deductivism. According to hypothetico-deductivism, theories are hypothesized and then experimental consequences deduced from them for testing (though the idea of a “method of hypotheses” occurs in Plato). Hence, though Bacon is often referred to as an inductivist, he certainly did not advocate a simplistic form of enumerative induction.³ Later philosophers of science who further elaborated inductive methodology include Mill and Whewell as well as many of the logical empiricists who did important work in the foundations of probability and inductive logic.

Karl Popper criticized all forms of inductivism whether sophisticated or not, and argued for falsificationism, which is the thesis that justification is solely about falsification and never involves induction. Popper thought that induction of all kinds was fallacious and argued that science proceeds by conjecture and refutation, and that theories are only ever so far unfalsified, never confirmed. This is part of his “critical rationalism” in epistemology that was based on his appreciation of the fact that in the history of science even theories that seemed incontrovertibly established by weight of positive evidence had later been shown to be erroneous. The best example is Newtonian mechanics, which Kant had sought to show to be knowable certain and known a priori. Newton’s three laws plus the force law for gravitation described phenomena in diverse domains and to very high accuracy. Yet they had been overthrown by the theories of relativity. This fact led Popper to adopt a radical form of

fallibilism according to which no part of our theories was to be regarded as inviolable; he thought we should actively try to show our most cherished theories to be false as this is how science progresses. According to him the best theories are those that make precise quantitative and counterintuitive predictions since they are most falsifiable. On the other hand, Popper was scathing about vague and general theories that could be modified to accommodate recalcitrant experience. However, his talk of “corroboration” of theories that survive many attempts to falsify them has led many philosophers of science to accuse him of equivocation, and in any case, most philosophers of science now accept that the inductive or probabilistic logic of positive support bestowed on theories by evidence, or confirmation, is central to scientific reasoning. However, core ideas of Popper, such as the emphasis on falsifiable statements and the avoiding of ad hoc modifications to theories to save them from refutation, remain very popular among scientists.⁴

Popper proposed his hypothetico-deductivism is the setting of a distinction between the contexts of discovery and of justification in science.⁵ This was at odds with the conception of scientific method due to Bacon and Descartes (and many subsequent philosophers), who claimed to offer (very different) rules that would lead us from ignorance to knowledge. Popper argued that scientific theories arise by all manner of means and cannot be “deduced from the phenomena” as Newton claimed his laws were. According to Popper, the origins of a theory in the thought processes of individuals and groups are a matter for psychology and/or sociology not logic or philosophy of science. What matters is that once a hypothesis is arrived at, deductions of specific and precise empirical consequences are made allowing it to be rigorously tested. Many other philosophers and scientists separate questions concerning the evidence for a theory and its rational acceptability from the contingencies surrounding who first proposed a theory. For example, several people independently thought of the basic idea of quarks, and the Higgs boson owes its names to Peter Higgs though others had similar models. Nonetheless, the extent to which there is a context of justification independent of psychological and sociological facts is very controversial following Kuhn’s work, as discussed below.

The “logic” of science is still contested among those who accept that some form of induction based on experience is at the heart of the scientific method, and a great deal of work studies and attempts to formalize the confirmation relation, and to provide measures of the degree of quantitative support between evidence and theory using the mathematical theory of probability. The most basic idea about the relation between theory and evidence is that a generalization is confirmed by observations of its instances. This is often known as Nicod’s criterion after the French philosopher of science who advocated it. Carl Hempel’s classic paper on the paradoxes of confirmation raises

various problems that arise when Nicod's criterion is conjoined with other plausible assumptions.⁶ For example, the equivalence condition that states that the degree of confirmation between a theory and some evidence should be invariant under logical equivalence, so that, for example, if observation of a black raven confirms that all ravens are black, then so too should observation of a white swan, since it is a nonblack non-raven and so confirms all nonblack things are non-ravens. This is the Ravens Paradox and it together with others, including notably the Grue problem due to Nelson Goodman, led some philosophers of science to conclude that the confirmation relation is not purely logical and to advocate historical theories of confirmation that make which theory is most supported by the evidence explicitly depend on historical facts about the development of the theory.⁷ Hence, for example, Ellie Zahar argues that what made Special Relativity better than its rivals was that it was part of a productive research program that led to General Relativity and to search for symmetry principles in twentieth-century theoretical physics.⁸

Among those philosophers who think that the scientific method and theory choice can be understood in terms of an ahistorical logic of confirmation, most formalize it using probability theory. The probabilistic turn in philosophy of science in the twentieth century matches the statistical turn in science. Population genetics, epidemiology, and economics are among the sciences that drove this development, and the discovery that radioactivity appeared to be fundamentally random led many philosophers of science to think that there could be irreducible single-case chances in the world.⁹ The classical theory of probability is due to Pascal, Bernoulli, Huygens and Leibniz but a full mathematical theory of probability awaited the axioms due to Kolmogorov. However, these apply to measures that are not probabilities too. It is also contested whether the axiom of countable additivity holds for probabilities. Other fundamental questions about probability include: How do probabilities relate to actual relative frequencies of events in the world?¹⁰ Are there really single-case probabilities and if so how do they relate to generic probabilities?¹¹ Are all probabilities based on ignorance? Is objective chance compatible with determinism? Most important for the scientific method is the matter of how probability relates to rationality and inductive inference. Various paradoxes of probability including Bertrand's paradox and the lottery paradox are the subject of much recent work.

Probabilism is the name given to the view that beliefs really come in degrees (credences) rather than being all or nothing, and that accordingly that they can be quantified. Degrees of belief are often thought to be rational only if they satisfy the axioms of the probability calculus. The most famous argument for this view is the so-called synchronic Dutch Book argument. There is a good deal of controversy about whether or not this argument is successful, and even more about the so-called diachronic Dutch Book argument that is advocated

as showing that beliefs must be updated in the light of new evidence using so-called Bayesian conditionalization based on Bayes' Theorem. The latter states that the conditional probabilities relating two propositions are related as follows: $P(c/e) = P(c).P(e/c)/P(e)$, where $P(a/b) = P(a \& b)/P(b)$. In conditionalization $P(c)$ is called the prior probability of c and $P(c/e)$ is interpreted as the (posterior) probability that c should be assigned after e has been learned.

Bayesianism is a form of epistemology based on probabilism and conditionalization that has been hugely influential in the philosophy of science and among scientists.¹² Bayesians regard probabilities as personal or epistemic in the sense of being assignable not to the world but to agents based on their doxastic or epistemic states. However it is important not to conflate two dichotomies. The first is between subjective and objective probabilities, and the second between ontic and epistemic probabilities (where the latter means probabilities). Nobody thinks there are subjective ontic probabilities and many who regard probabilities as degrees of belief also think there are no true objective probabilities. This tradition originates with Frank Ramsey, Bruno de Finetti, and L. J. Savage and finds contemporary expression in the work of many decision theorists and so-called subjective Bayesians. On the other hand, some Bayesians, for example, John Maynard Keynes and E. T. Jaynes, argue that probabilities may still be objective even though they are epistemic. The fundamental idea is that probability should be shared equally among possible outcomes for finite sample spaces. This is known as the Principle of Indifference and is formulated by Keynes as saying that whenever there is no evidence favoring one possibility over another they have the same probability. Jaynes' (1968) mathematical version is based on information theory and called "the maximum entropy principle."

Those who believe in objective ontic probabilities often refer to them as "objective chances" and then tackle the problem of how they should relate to degrees of belief. David Lewis' Principal Principle has spawned a huge literature on this issue.¹³ Other work in Bayesianism epistemology relates to causal inference via so-called Bayes' nets,¹⁴ to measures of the coherence of hypotheses based on the relationship among probabilities,¹⁵ and to approaches that take conditional probabilities as primitive. There is also a sizable literature on classical statistical inference and the standard scientific statistical methods of significance testing and the calculation of confidence intervals.¹⁶

While Bayesians and others continue to analyze and explicate the scientific method and the rules of rationality, some philosophers have become profoundly skeptical about the rationality of science and deny that there is a scientific method as opposed to many methods used by different sciences. Thomas Kuhn's *The Structure of Scientific Revolutions* (and also Paul Feyerabend's *Against Method*) inspired many to hold these views and despite Kuhn later denying the charge that he had reduced science to mob psychology,

many have taken his work to show that scientific theory choice owes as least as much to individual, social, and political values and idiosyncrasies as it does to the evidence. Kuhn's work challenged the received view according to which science is cumulative and progressive, epitomizes rationality in virtue of theory choice having an underlying logic that is free of social and political values, and thus can be sharply distinguished from other kinds of belief systems. Many sociologists and historians of science now study theory development and change in science as if they were studying any other belief system, and their methodology is sometimes radically at odds with the rational reconstructions of theory and evidence that are the currency of analytic philosophy of science. Likewise many of those who work in science and technology studies reject the traditional project of showing why a theory was chosen in the light of the way the world is as Whiggish and fanciful. Kuhn's influence on the philosophy, history, and sociology of science and the wider culture has been extraordinary.¹⁷ The relationship between philosophy of science and the history of science remains contested.

Scientific theories are supposed to be based on *observation*. However, since Popper and Kuhn, many philosophers of science have pointed out that observations are guided by and described in terms of theories raising the question of whether or not they can provide a neutral foundation for scientific knowledge. The logical empiricists, notably Carnap, often divided the language of science into observational terms and theoretical terms, in the cause of explicating the meaning of the latter in terms of the former. This view is no longer taken seriously and furthermore it is questioned whether we can distinguish between observational data and our interpretations of it? The observation language being theory-independent was supposed to represent the fact that observation provided a neutral foundation for the testing of scientific theories and scientific knowledge. Kuhn, along with N. R. Hanson and Paul Feyerabend, argues that observation is theory-laden. What exactly this means and whether it is true is still much-discussed following a debate between Paul Churchland and Jerry Fodor.¹⁸ In any case, it is now recognized that most scientific terms including observation terms do not have fixed and precise meanings, and that observation now almost always involves the use of sophisticated measuring devices and the recording of magnitudes that are unobservable and that are often only arrived at by machine-assisted computation.

Kuhn's work was published in an encyclopedia of the unified sciences. The logical empiricists explicitly advocated the unity of science but the latter turns out to mean many things. Putatively and primarily there is the epistemological unity provided by a single scientific method. Kuhn suggests that science is not unified in the sense of being based on a single set of fundamental methods and aims for all the sciences at a time, and that conceptions of the right methods and aims change when paradigms change. Since, Kuhn and others work

on the unity of science is much disputed and many argue that there are many scientific methods and quite distinct ones in fundamental physics on the one hand, and the social sciences on the other. Nancy Cartwright argues that science is an overlapping patchwork of models without the kind of hierarchy supposed by reductionists and physicalists, and without even consistency between different parts of the whole. On the other hand, many remain committed to strong or weak forms of reductionism and argue that consistency and unification are central to science and its methods (see below for the metaphysics of reduction and unity).

The problem of underdetermination of theory by data remains the subject of much interest. It is seen as arising from the related Duhem problem that is ever-present in real science, as recently with the apparent evidence of particles travelling faster than light. The Duhem problem is that whenever a theoretical prediction conflicts with the evidence it is never a matter of straightforward deduction where to localize the falsification among the host of hypotheses and assumptions from which the prediction was derived. This latitude means that scientists faced with an apparent problem for fundamental theory have to choose whether to revise the latter, or whether to tinker instead with one or more of the rest. Kuhn argues that there is an essential tension between conservative and revolutionary tendencies in science and that how to balance them cannot be prescribed by any logic of the scientific method.¹⁹ Another fundamental tension is that between avoiding two kinds of error in science namely false positives and false negatives. Suppose that a patient is being tested for a disease; a false positive is when the test says he or she has it when they do not, and a false negative is when the test says he or she is clear when in fact they have the disease. Avoiding both kinds of error is obviously important, but the more one avoids one of them the more one is likely to fall into the other. When scientists are making inferences, for example, about whether some experimental results amount to the discovery of a new particle, they must weigh the goal of avoiding leaping to a mistaken conclusion against the goal of not missing out on important truths.

It has long been thought that simplicity is a theoretical virtue that may help solve the problem of theory choice. Okham's razor famously prescribes against hypothesizing more entities than are needed to account for the phenomena. This is not as simple as it sounds and there are qualitative and quantitative notions of simplicity. Much work has concerned formalizing the notion of simplicity in the context of the problem of fitting a curve to a set of data points.²⁰

Logical positivism and logical empiricism had a profound influence on the whole of philosophy. Their disdain for metaphysics was matched by their enthusiasm for science, mathematics, and logic. Mathematical logic was seen as a tool for the clarification of scientific concepts and the settling of all epistemological questions. The basic framework they used was first-order logic,

and the nonlogical vocabulary was usually partitioned into observational and theoretical terms. For some time, it was hoped that the meaning of the latter would be entirely characterized in terms of the former plus logical structure. According to the so-called “received view” that is the *logical empiricist* model of science (most developed by Carnap), theories are essentially linguistic entities consisting of a syntax of theoretical and observational terms, and containing *correspondence rules* linking theoretical terms with the results of observations.²¹

Philosophers have sometimes imagined scientific theories to make contact with the world, by sets of initial conditions being conjoined with the fundamental laws, and then specific details being deduced. It is now universally recognized that the application of fundamental theories involves a creative process of model construction that is often multistage and involves different forms of abstraction, approximation, and idealization. Modeling is the generic term for this activity and it often now involves computational models and simulations, as well as requiring computations for calculations. Mary Hesse’s seminal work on models and analogies in science is now accompanied by a wealth of literature on models in science and specific discussion of causal modeling and simulations.²²

It was once believed that science could deliver the complete and exact truth but the history of science has taught us that approximate truth is all for which we ought reasonably to hope. However, science can be very accurate indeed while still being strictly in error. For example, Newtonian mechanics and gravitation is accurate to one part in a million in its description of planetary orbits in our solar system. Nonetheless, the Newtonian account of the nature of space and time, and radiation, matter, and forces is quite wrong about fundamentals. Formalizing what it means to say a theory is approximately true turns out to be very difficult. Popper famously proposed a simple definition in terms of sets of consequences that was quickly shown to be inoperable. Since then a variety of models have been proposed. Much of this work connects with the literature on approximation in general.²³

2 The Metaphysics of Science

While many of the founders of the philosophy of science were hostile to metaphysics, Popper and his school always acknowledged the importance of the latter in the history of science. According to Lakatos, metaphysical hypotheses could form part of the “hard core” of research programs. As such he thought with Popper that metaphysical hypotheses were not falsifiable, but nonetheless perfectly meaningful and often useful in providing heuristics for the development of theories. Eventually, a research program may stop bearing

fruit and when it does its metaphysical framework will be abandoned; so in this sense metaphysics, on their view, is hostage to empirical fortune. For example, Cartesian metaphysics posits a plenum of purely mechanical matter was of heuristic use leading to Descartes' model of the formation of the solar system for instance, and incorporated a ban on action at a distance and an emphasis on geometrically representable motions that was productive for scientific theorizing. However, the success of Newton's theory of gravitation and its incompatibility with Cartesianism led to the latter's demise. The heuristic to avoid positing action at a distance survived it and guided the development of field theory and the eventual positing of local action and finite velocity propagation of gravity.²⁴

Contemporary philosophy of science intersects with the analytic metaphysics of causation, dispositions, laws, and modality. This is an important development in the philosophy of science since the mid-twentieth century and mention is often made of the "metaphysics of science." Under this heading there is much recent discussion of scientific examples of dispositions, essences, individuals, kinds, and powers, such as conductivity, atomic number, species, species, and force.²⁵ It is no accident that the rise of the metaphysics of science accompanied the rise of scientific realism after the demise of logical empiricist forms of antirealism (see below), because it is only if science is taken as telling us about more than the phenomena that metaphysics and science overlap. Early work on the metaphysics of science arose out of the debate about theories of confirmation and reference and the seminal discussion of natural kinds by Quine and Putnam.²⁶

Traditional issues in the metaphysics of causation, such as the relationship between causation and regularities, between causal relations and necessary and sufficient conditions, and between causation and counterfactuals, are much discussed in recent philosophy of science. Philosophers of science have also been much exercised by the relationship between causation and explanation (see below).²⁷ Many scientific theories give causal explanations based on statistical relations, and some philosophers now believe that physics tells us that the world is fundamentally indeterministic. Hence, the traditional view of a cause as being sufficient for its effect may have to be abandoned and replaced by the idea that causes merely alter the probabilities of their effects.²⁸

Many philosophers regard laws of nature as central to science. Hence, one reason why some have argued that the social sciences are not genuinely scientific is the alleged dearth of laws of economics and society. Other ideas associated with laws include those of generalization, regularity, pattern, stable relationship, symmetry, and invariance. The latter two notions are of great importance in the philosophy of physics. Different kinds of laws include laws of motion or state evolution over time, laws of coexistence, conservation laws, laws as principles, phenomenological versus fundamental laws, deterministic

versus probabilistic laws. There is also the matter of the relationship between laws and whether laws in high-level sciences can be derived from those in more fundamental ones.

For the logical empiricists, laws of nature are general statements that figured as axioms in theories construed as sets of sentences generated by inference from those laws. If boundary conditions are included predictions of particular events become possible. For Carnap, the statement of precise laws is an essential characteristic of science. The classic discussion of laws in Hempel's *Philosophy of Natural Science* concerns their role in explanation (see also Ernst Nagel's *The Structure of Science*). Recent discussions of laws of nature has focused on the metaphysics of laws.²⁹ What is a law of nature and do they differ from generalizations that happen to be universally true, like no gold spheres are bigger than the earth? (No plutonium spheres are bigger than the earth is an analogous generalization that seems to be law-like.)

Empiricists usually deny the existence of natural necessity (see Hume on causation and induction) and hence any inherent difference between lawlike and accidental regularities. The simple or naïve regularity theory of laws says it is a law that all As are Bs iff all As are Bs. The main problems for this regularity theory seem to be that on the one hand not all regularities are laws, especially single-case regularities and disjunctions of them, regularities with disjunctive or grue predicates, and vacuous regularities; and on the other hand not all laws are regularities because of unrealized physical possibilities, uninstantiated laws, and regularities that seem purely accidental. The sophisticated regularity theorist therefore places restrictions on what regularities are to be counted as laws. A. J. Ayer and Richard Braithwaite both argued that it is our cognitive attitudes that determine which regularities are laws. The main problem with this is that laws can be unknown. Which of the unknown regularities are laws and which are not can only be a matter of what our attitude to them *would* be if we knew them. This is obviously problematic since such counterfactuals would seem to rely upon laws themselves. Why do we have different attitudes to different regularities? They are either arbitrary or grounded in some objective difference between them. If the former then no good, if the latter then some substitute for natural necessity must distinguish laws from accidents.

An important view of laws, defended by Frank Ramsey and David Lewis and originating in Mill (the Mill-Ramsey-Lewis regularity theory of laws) is that laws are the "consequences of those propositions which we should take as axioms if we knew everything and organised it as simply as possible in a deductive system" (Ramsey pp. 128–32). Lewis' take on this is that laws are the result of a trade-off between simplicity and strength. Laws are the theorems and axioms of deductive systems that achieve the best combination of simplicity and strength.³⁰ The problems with this view include: simplicity may not be

an objective notion nor strength, what achieves the best balance of strength and simplicity may not be agreed upon by all—rationalists might prefer simplicity whereas empiricists might prefer strength, we might get more systematic laws using grue or disjunctive predicates, and there could be equally systematic but different sets of laws (see Coherentism in epistemology).

On the other hand necessitarians, notably David Armstrong, Fred Dretske, and Michael Tooley, hold that laws of nature are relations between universals.³¹ They regard laws as part of modal metaphysics and are realists about nomic necessity. On such views, Laws imply universal truths but universal truths do not imply laws (because laws generate opaque contexts). This approach seems to offer an account of how laws support counterfactual statements, and is better placed to deal with the relation between laws, explanation, and inference. However, it has many costs other than the ontological commitment to universals. The nature of the necessitation relation between universals remains mysterious and how can we make sense of the inference from “F-ness \rightarrow G-ness” and “this is F” to “this must also be G” if the laws of nature are themselves contingent? There is also the matter of accounting for probabilistic laws.

Some philosophers of science are skeptical about the traditional views of the importance of fundamental laws in the analysis of scientific theories including Bas van Fraassen, in his *Laws and Symmetry*, part I, and Nancy Cartwright, in her *How the Laws of Physics Lie*. In the latter Cartwright made a distinction between phenomenological laws and fundamental laws, and argued that the latter are not true of the actual world. (Cartwright also argues for an ontology of causal powers and against Humean accounts of causation in science.) According to her analysis of laws in physics putatively fundamental laws are not true because they are always really *ceteris paribus* laws. *Ceteris paribus* laws are laws that only hold in the absence of confounding factors. For example, the law of supply and demand in economics is often thought of as a *ceteris paribus* law because in actual economic systems there is no linear relationship between supply and demand because of the presence of monopolies, cartels, and so on. Many special science laws seem to be *ceteris paribus* laws and the problem is how to spell out what is meant by all things being equal without reducing them to vacuity.³²

Science is now vast and hugely integrated across many domains and levels of description. For example, molecular biology integrates physics and chemistry with the study of life, neuroscience studies computational processes that involve chemical signaling as well as electromagnetic phenomena, and evolutionary game theory is employed in economics and biology. However, what does the unity of science amount to and what are its metaphysical implications? Can biology and chemistry be *reduced* to physics? What would reduction consist of if it were possible? These issues have important implications in the philosophy of mind, where the question of whether or not the mental

is reducible to or *supervenes* upon the physical is of great importance, and in ontology, since we need not include entities that are reducible in our ontology of basic objects. They are the subject of much classic and recent work. The discussions of reduction by Nagel and Hempel form the foundation of this subject along with earlier debates about emergentism in the twentieth century and discussions in the previous century by Mill. In the history of science there are many great achievements of unification and reduction. Foremost is Newton's mechanics with a law of gravitation that unified the terrestrial and extraterrestrial domains. Notable others include Faraday and Maxwell's unification of electricity and magnetism, Helmholtz's work on energy in living systems that led to the conservation of energy, Pasteur's work on fermentation showing that chemistry could be applied to biological processes, the genetic basis of inheritance in the DNA molecule, and the periodic classification of the elements and the common system of measurement for scientific units.

Cartwright and John Dupre are well-known critics of the idea that science is unified. Cartwright argues that the relationship between theories, models, and reality undermines what she calls "fundamentalism" about laws. Dupre argues for what he calls "promiscuous realism." Such a view is often defended in philosophy of biology where it frequently seems as if a single scientific term such as species or gene is understood differently in different parts of the science, and for which pluralism about meaning and reference is proposed.³³

3 The Epistemology of Science

There is obviously a substantial overlap between the epistemology of science and the theory of the scientific method and many of the topics discussed above have relevance for what follows. In this section I focus on epistemological problems that sometimes divide those who agree about many other aspects of scientific methodology. The most fundamental division in the epistemology of science is probably that between realists and antirealists. However, there are many different forms of realism and antirealism and many subtleties to the debates between them. Roughly speaking though, realists emphasize that science gives us knowledge of an unobservable world to explain observable phenomena, while antirealists emphasize the empirical adequacy and instrumental power of science and are skeptical about the metaphysics of causation, essence, and law that often accompanies belief in unobservable entities. This conflict has its roots in antiquity and the mediaeval disputes about universals, and in early modern philosophy; Descartes and Locke were pioneers of realist philosophies of the material world, while Berkeley, Mach, and Duhem criticize realism. There is a close link between the fortunes of scientific realism and that of atomism in physics and chemistry, and scientific realism became the

dominant view among philosophers of science in the mid-twentieth century. Whether or not one adopts scientific realism in general, it is possible to take an antirealist view of particular theories.³⁴

The standard form of scientific realism came to prominence in the aftermath of logical empiricism and is associated with the early work of Hilary Putnam, and the writings of Richard Boyd, Paul Churchland, and Ernan McMullin. Well known contemporary realists include Brian Ellis, David Papineau, and Stathis Psillos.³⁵ Realism became defined by several components as it was articulated in contradistinction to the forms of antirealism that were prevalent in the early twentieth century. The previous generation of antirealists such as Duhem, Mach, and Poincare did not advocate antirealism as a mere interpretation of scientific theories, but rather as part of their ideas about the methodology of science, in particular, they were all united in opposing the atomist research program. When atomism triumphed and unobservables such as radio waves and cathode rays became essential tools of technology, antirealists tended not to argue that theoretical terms such as “atom” should not be employed within science. Contemporary science is so replete with theoretical terms and laws stated in terms of them that the idea that we should divest with them is no longer regarded as worthy of serious consideration. Instead, there were various proposals for ways of rendering them metaphysically innocuous. Two such principal proposals were:

semantic instrumentalism: the theoretical terms of scientific theories should not be taken literally as referring to unobservable entities, because they are merely logical constructs used as tools for systematising relations between phenomena. Theoretical hypotheses are not assertoric.

reductive empiricism: Theoretical terms can be defined in terms of observational concepts, hence statements involving them are assertoric. Scientific theories should not be taken literally as referring to unobservable objects. Their theoretical terms are used in disguised ways of referring to observables.

It became widely agreed that both these proposals fail and that theoretical terms should be taken literally. It was also appreciated that theoretical terms per se were not the issue since unobservables could be referred to with ordinary terms as in “tiny particles too small to see.”³⁶ Recent debate has accepted the semantic presupposition of realists, namely that theoretical discourse about unobservable entities, properties, and processes is meaningful and to be taken literally, irreducibly, and truth-aptness and as describing an unobservable world of novel objects and properties. Those dissenting from scientific realism now tend to attack either its metaphysical or its epistemic components. The former

is the standard metaphysical doctrine of an external world whose nature and state is (largely) independent of our beliefs and desires about it and the structure of our cognition. The latter is the claim that our best scientific theories really do refer to unobservable entities and processes and are approximately true and known to be so. The scientific realism debate is largely though not entirely about the epistemological question as to whether we can know that our theories are true given that they are taken at face value as talking about, for example, quarks and electrons.

Van Fraassen's "constructive empiricism" is the most well-known form of antirealism. While attacking the positive epistemology of scientific realists and especially the idea that scientific realism is the only rational option, he defines his own view not in epistemological terms but in terms of the aim of science. As mentioned above it has long been controversial exactly what it is. Everyone agrees that there are proximate aims such as to model particular systems or to predict or explain certain phenomena. Individual scientists may of course also have all sorts of aims that have nothing to do with science, or their own particular obsessions about it. However, what is the ultimate aim if any of science as such. Mountaineering may have many aims including enjoying the great outdoors and developing skills and endurance, but presumably the fundamental aim is to reach a summit. Does science have such an intrinsic telos? Some standards answers are as follows: (a) the aim of science is to produce knowledge in the sense of justified or reliable beliefs; (b) the aim of science is to deliver true beliefs; (c) the aim of science is the manipulation of the world for the furtherance of human ends; and (d) the aim of science is to explain the phenomena and uncover the hidden workings of the world and its laws. (d) is clearly compatible with (a) and (b). (c) is an expression of pragmatists who have always, like instrumentalists, emphasized the usefulness of description over the satisfaction of explanation.

Van Fraassen argues that the aim of science is to produce true beliefs, but only about the phenomena not about the unobservable world. He articulates this view in expression of the broader empiricism that led earlier generations of philosophers of science to eliminativism or instrumentalism about theories of the unobservable world. While van Fraassen is usually taken to be proposing an epistemology of science he in fact confines himself to axiology and his constructive empiricism is compatible with belief in unobservable entities. His arguments against scientific realism are focused on explanation and inference and the metaphysics he thinks goes with them.³⁷

Many philosophers of science advocate a central role for inference to the best explanation in the epistemology of science. It is commonly argued that there is no special problem with inference to unobservable entities, and hence with scientific realism, because inference to the best explanation (IBE) is an essential part of ordinary induction. This is often called the "explanationist"

defense of scientific realism. It is argued by philosophers such as Richard Boyd, David Papineau, and Stathis Psillos, that scientific realism is a naturalistic theory of science, and that the best explanation of the success of science is that it is correctly describing the unobservable causes of what we notice. Such an explanation may be offered at the level of specific theories or of the whole of science. Accordingly there is a distinction made between local and global, or “retail” and “wholesale,” applications of Inference to the Best Explanation (IBE) and other arguments to scientific theories.³⁸

The no-miracles argument may be nothing more than an intuition pump but it continues to motivate many philosophers and scientists to be scientific realists. There are various versions and it is sometimes known as the “ultimate argument” for scientific realism. All are premised on the success of science, but it is not agreed what kind of success. There are stronger and weaker kinds of empirical success. Predictions may be more or less accurate, and they may be more or less unexpected. A successful prediction of a full Moon is less impressive than the successful prediction of the return of a comet or the occurrence of an eclipse. In addition to the prediction of familiar events and rare events there is the prediction of completely unanticipated forms of phenomena. Since Popper celebrated the prediction of the bending of light by the Sun by General Relativity it has been common to cite novel predictions

The notorious underdetermination problem mentioned above is often seen as a particular problem for scientific realism and the third chapter of *The Scientific Image* is devoted to it. However, many have argued that the problem of induction is effectively a form of underdetermination and that theoretical virtues such as explanatory power must be deployed to overcome it and hence can also be used to overcome the underdetermination of theoretical hypotheses over and above the choice of a class of empirically equivalent ones. The debate about the problem of underdetermination has reached something of a stalemate, though it has been recently invigorated by Kyle Stanford’s advocacy of what he claims is a new form of underdetermination known as underconsideration.³⁹ Many philosophers have regarded it as akin to the problem of external world and hence as a kind of skepticism that while irrefutable is also uninteresting. The most compelling arguments against scientific realism are those from theory-change in the history of science, and the most discussed is the pessimistic meta-induction:

- (a) There have been many empirically successful theories in the history of science that have subsequently been rejected and whose theoretical terms do not refer according to our best current theories.
- (b) Our best current theories are no different in kind from those discarded theories and so we have no reason to think they will not ultimately be replaced as well.

- (c) So, by induction we have positive reason to expect that our best current theories will be replaced by new theories according to which some of the central theoretical terms of our best current theories do not refer, and hence we should not believe in the approximate truth or the successful reference of the theoretical terms of our best current theories.

Realists usually respond to this argument by restricting realism to theories with some further properties (usually, maturity and novel predictive success) so as to cut down the inductive base (see Psillos 1996). Peter Lewis (2001), Marc Lange (2002), and Magnus and Callender (2004) regard the pessimistic meta-induction as a fallacy of probabilistic reasoning.⁴⁰ However, there are arguments from theory change that are not probabilistic based on cases of mature theories that enjoyed novel predictive success, notably the ether theory of light and the caloric theory of heat. If their central theoretical terms do not refer, the realist's claim that approximate truth explains empirical success will no longer be enough to establish realism. There are two common (not necessarily exclusive) responses to this:

- (I) Develop an account of reference according to which the abandoned theoretical terms are regarded as successfully referring after all. Realists developed causal theories of reference to account for continuity of reference for terms like "atom" or "electron," even though the theories about atoms and electrons have undergone significant changes. The difference with the terms "ether" and "caloric" is that they are no longer used in modern science. However, as C. L. Hardin and Alexander Rosenberg (1982) argue, the causal theory of reference may be used to defend the claim that terms like "ether" referred to whatever causes the phenomena responsible for the terms' introduction.⁴¹ This is criticized by Laudan (1984).
- (II) Restrict realism to those parts of theories that play an essential role in the derivation of subsequently observed (novel) predictions, and then argue that the terms of past theories which are now regarded as non-referring were non-essential and hence that there is no reason to deny that the essential terms in current theories will be retained. (Kitcher 1993)⁴²

The debate about standard scientific realism continues but other forms of realism have also been the subject of much discussion in recent years. Entity realism, is defended by Ian Hacking and Nancy Cartwright.⁴³ They argue in line with a more general pragmatism that realism should be based on our causal interaction with unobservable objects not on high theory and fundamental laws. On the other hand, structural realism was revived by John Worrall as a form of realism that denies that we have knowledge of unobservable entities

and celebrates instead of the structural knowledge expressed in mathematical relationships, which are retained even when the ontology of theories changes significantly. Ontic structural realism was proposed by Ladyman (1998) who built on Worrall's ideas but also intended to account for various metaphysical aspects of contemporary physics. It has since been developed in various guises by Steve French and James Ladyman and their collaborators and others.⁴⁴

Realists often cite novel predictive success as the main motivation for scientific realism. This follows the distinction that is often made between accommodation, the theoretical recovery of a known fact, and prediction, the derivation in advance of something subsequently observed. There has been a series of recent contributions to the debate about predictivism, which holds roughly, that there is more confirmation, or perhaps, only confirmation of a theory by evidence, if the latter was not.⁴⁵

What does it mean to give a scientific explanation of something? Presumably such an explanation must involve a scientific theory somehow but can we give necessary and sufficient conditions for something to be an explanation? The orthodox view of explanation was the "covering-law model." It came in two forms: the deductive-nomological and the inductive-statistical. Both are now widely contested and various alternative accounts of explanation have been proposed. Important questions that such accounts must answer include: do generalizations in explanations need to be laws of nature, and are all proper explanations causal?⁴⁶

While many scientific explanations directly appeal only to the causal powers of concrete objects, some seem to make reference to more or less abstract facts expressed in the language of mathematics. In the advanced sciences mathematics is indispensable and this is the basis of an argument by Quine and Putnam to argue for realism about mathematics.⁴⁷

4 The Philosophy of the Sciences

Terminology often distinguishes between physics and the special sciences, however, areas of the former such as optics, or fluid mechanics, are really kinds of the latter, so it is better to distinguish between (putatively) fundamental physics and the special sciences. Many of the philosophical problems associated with the latter have to do with the status of a particular special science in relation to a more fundamental one. For example, questions in the philosophy of psychology often have to do with how mental states and processes related to neurophysiological ones.

Cognitive science may be construed as embracing artificial intelligence, linguistics, neuroscience, psychology, and behavioral science and the relevant philosophical issues in the philosophy of mind and language are the subject

of much debate among scientists. Recent work on neuro-imaging raises new epistemological issues and applications of traditional debates about the inference from observation to theory. Longstanding debates in the philosophy of perception about direct and indirect realism and the objects of experience are now tackled in the light of the latest findings of neuroscience. Fundamental questions about consciousness, free will, and qualia remain and there is extensive philosophical and scientific work on them, and also on the nature of the emotions and moral psychology. Neuroimaging and sophisticated experimentation can be combined to reveal how the different systems of the brain affect our behavior. The ancient debate about the *a priori* has a contemporary form in the discussion of nativism about language and concepts in developmental psychology. The status of explanations and theories in evolutionary psychology is hotly disputed, as is the degree to which the mind should be thought of as modular rather than general purpose. The problem of the relationship between folk psychology and an envisaged reductive science of the mind raises issues about the nature of theories, and, in particular, whether folk psychology should be thought of as a proto-scientific theory. Furthermore, theories of mental simulation based on the discovery of mirror neurons have transformed the way questions about knowledge of other minds are asked. Aesthetics may be informed by detailed knowledge of how aesthetic judgment functions.⁴⁸

Philosophy of physics is largely concerned with the study of spacetime physics, quantum physics, thermodynamics and statistical mechanics, and cosmology. Recent developments have seen more work on classical physics, especially Lagrangian and Hamiltonian mechanics, chaos theory, and electromagnetism, and an increased interest in the physics of phase transitions and quantum field theory and quantum gravity. The philosophy of cosmology is an important new subdiscipline.⁴⁹

Philosophy of biology has grown extensively and now has specialist journals devoted to it though much relevant work is in theoretical biology journals. There are prominent debates about the nature of causation in evolutionary biology and about the definitions of fitness and function, as well as traditional philosophical questions about individuals and natural kinds applied to species and organisms. The levels of selection problem concerns whether evolution acts directly on genes, organisms, and/or groups and is the subject of much discussion among biologists. Group selection may be part of a naturalistic ethical framework since it is sometimes thought to explain the existence of altruism.⁵⁰

Evolutionary game theory is used both in biology to understand optimization by natural selection, mating strategies, and how organisms solve problems of coordination and competition in collective pursuits such as foraging or defense. It also finds application in economics and social science. Recent

work has considered analogies between adaptation and the maximization of expected utility: that is, between Darwinian natural selection and rationality where each are thought of as requiring a certain kind of optimization.⁵¹

The clinical and biomedical sciences account for a huge proportion of scientific funding and research and raise philosophical issues that do not arise in other areas of science. Questions about what it means to live a flourishing human life are fundamental to medical ethics and in the philosophy of medicine issues such as the nature of disease and the taxonomy of human psychopathology are thoroughly entwined with philosophical dispute about the human condition and the meaning and purpose of life. There are also debates about evidence-based medicine and whether or not randomized controlled trials are absolutely required to test drugs and other treatments.

Computation raises many philosophical issues. The most fundamental is what is a computation? The answer that is usually given is that it is anything that can be done by a universal Turing machine. The physical Church-Turing thesis states that any computation that can be physically carried out is computable by a Turing machine. However, it is controversial as is the status of natural computation and whether or not information-theoretic and computational accounts of objective reality are coherent. There is lively discussion in philosophy of science about the nature of computation, and about the identity conditions and status of algorithms, programs, and data. There are close connections with debates in philosophy of mind and psychology about computational theories of mind and the naturalization of representation.

Furthermore, computation is about to play a new role in science. Previously, computation has not been used for theorizing but only for computation, modeling, and simulation; however, data mining and pattern recognition are now so sophisticated that it is routine for machines to learn to find projectible regularities in data for themselves. This may lead to the automation of science and reinvigorate old debates about the logic of discovery. Already many philosophers of science have contributed to theories of machine and formal learning and the intersection of these with Bayesianism, belief revision, and decision theory. The debates about models and modeling have been enriched by new work on computer simulations and the role of computer models and their predictions in controversy about anthropogenic climate change.⁵²

Philosophy of chemistry is another area of recent development and is now a flourishing subdiscipline with fascinating connections to the philosophy of physics and to central philosophical issues about causation, emergence, ontology, and reduction that are usually discussed in the philosophy of mind. In respect of the former, some defend downward causation in the context of explaining molecular shape. This means strong emergence and the autonomy of chemical theorizing. There are debates about the existence of atomic orbitals, the extent of successful reduction in quantum chemistry, and how to

characterize chemical kinds. Furthermore, the historiography of chemistry is of crucial importance since the chemical revolution was a central example for Kuhn, and questions about theory choice, realism, and pluralism are much discussed.⁵³

Forensic science raises interesting examples for philosophers of science. The citation of genetic data and tests in criminal trials, like that of fingerprint evidence, involves explicit judgments about probability. Forensic evidence is regarded as scientific and so impartial and objective, but it is known that individuals performing scientific tests are prone to psychological biases such as confirmation bias, which is an important concern in other areas such as identification evidence. There is a large legal literature on these matters and on the so-called prosecutor's and defense attorney's fallacies, which are instances of the well-known base-rate fallacy.

The historical sciences such as paleontology raise special issues since there is no possibility of conducting experiments to rerun history.⁵⁴ The so-called complexity sciences are in fact diverse parts of the natural and social sciences that have in common the fact that they treat systems with a very large number of components interacting many times in such a way as to give rise to emergent phenomena. Complex systems are studied using approximate and statistical computational models. They raise issues about reduction and emergence and the epistemology of models and simulations. The use of information theoretic techniques is widespread in complexity science.⁵⁵

I have said virtually nothing about the special issues raised by the social sciences and the relationship between economics and decision theory and the philosophical study of rationality. Other important topics that fall within philosophy of science but which are also of much wider concern include the science of gender and race, scientific authority and decision making, and science funding, the social dynamics of science and the ways in which science is changing through information technology. There is also intense debate about the relationship between science and religion.⁵⁶

Notes

- 1 There are many excellent introductions to the philosophy of science including Hacking 1983, Brown 1993, Bird 1998, and Ladyman 2002. Anthologies of seminal papers include, Boyd et al. 1991, Papineau, ed., 1996, Curd and Cover 1998, Balashov and Rosenberg 2002, and Bird and Ladyman 2012. The key specialist journals for general philosophy of science are *The British Journal for the Philosophy of Science*, *Philosophy of Science*, and *Studies in History and Philosophy of Science*, together with the newer *European Journal for the Philosophy of Science* and *International Studies in Philosophy of Science*. Much philosophy of science also appears in *Erkenntnis*, *Synthese*, and general philosophy journals. There are excellent entries on many topics on the philosophy of science in the *Stanford Encyclopaedia of Philosophy*.

- 2 It is common for philosophers of science to apply theories of the scientific method to case studies and also for them to start with the latter and seek to generalize lessons about the scientific method. A recent collection on the demarcation between science and pseudoscience is: Pigliucci and Boudry 2013.
- 3 For Francis Bacon apart from his own works see chapter 3 of Gower 1997, chapter 2 of Woolhouse 1988, and Urbach 1987. Gower's book is an excellent guide to theories about the scientific method and has chapters on many of the other philosophers of science discussed below.
- 4 The first chapter of Popper's *Conjectures and Refutations* is a clear introduction to his views. Modifications to theories are said to be ad hoc when they accommodate data but do not lead to new predictions. Lakatos identified several more specific ideas of ad hocness in Popper's work in his "The Methodology of Scientific Research Programmes" in his *Criticism and the Growth of Knowledge*, which contains many other classic papers delivered at a symposium about Kuhn's ideas (see below).
- 5 The distinction is usually attributed to Hans Reichenbach the logical empiricist; however, he did not understand it in the same way as Popper (see Glymour and Eberhardt 2011).
- 6 Hempel 1945, pp. 1–26 in his *Aspects of Scientific Explanation*, pp. 3–46. See Glymour 1980, Swinburne 1971, pp. 318–30, Earman and Salmon 1992, pp. 42–103, and chapter 2 of Brown 1993. A more recent discussion with references to much more current work on the theory of confirmation is Fitelson 2006, pp. 95–113.
- 7 Goodman 1955 and Musgrave 1974, pp. 1–23. The ravens paradox and the grue problem are also often taken to show that some theory of natural kinds and corresponding natural predicates must be posited to explain why we do not regard the claim that all emeralds are grue as confirmed by our evidence (see below).
- 8 See Zahar 1989, based on his paper "Why Did Einstein's Programme Supersede Lorentz's?" in two parts in *The British Journal for the Philosophy of Science*, pp. 95–123 and 223–62.
- 9 The history of probability in early modern science is described in Hacking 1975, and subsequent developments in statistical science are recounted in Hacking 1990.
- 10 Frequency theories reduce probabilities to finite relative frequencies. They face problems with single-case probabilities, and with the famous reference class problem (see Hájek 2012, which is an excellent introduction to the philosophy of probability including the other questions mentioned here).
- 11 Karl Popper's (1959) propensity interpretation of probability takes them as primitive and ontic but is not popular.
- 12 The classic works advocating a Bayesian approach to philosophy of science are Horwich 1982, Howson and Urbach 1989, now in its third edition, and Earman 1992. Critiques of Bayesianism include Glymour's "Why I am not a Bayesian," in his *Theory and Evidence*, and Papineau, ed., 1996, and Mayo 1996.
- 13 Meacham 2010, pp. 407–31.
- 14 Bovens and Hartmann 2003.
- 15 Douven and Meijs 2007, pp. 405–25.
- 16 Mayo 1996.
- 17 For Kuhn's ideas see his *The Structure of Scientific Revolutions* and *The Essential Tension: Selected Studies in Scientific Tradition and Change*, which expresses a more nuanced view. See also Bird 2000, and Ladyman 2002, ch. 4. See also Laudan 1977. Feminist and other critiques of science have drawn attention to cases in the history of science in which erroneous theories have been orthodoxy among scientists despite lacking evidential support because they fit with social and political values concerning gender, race, or certain groups of people. See, for example, Pinnick

- 1994, pp. 646–57 and Harding and Hintikka 2003. See also section 2 of Ladyman and Bird 2012.
- 18 Hanson 1958, ch. 1, Churchland 1979 and 1990, ch. 12, Fodor 1984, pp. 23–43, 1991, pp. 201–20, and in his *A Theory of Content and Other Essays*. Hacking 1983, chs 10 and 11. Shapere 1982, pp. 485–525.
- 19 There is an extensive literature on the underdetermination problem. It is often related to Duhem’s problem (Duhem 1914, ch. 6) of localized confirmation or falsification of theories within the mass of laws, background, and auxiliary assumptions and initial conditions needed to derive a prediction. For Quine on underdetermination see his “Two Dogmas of Empiricism,” in his *From a Logical Point of View*, and much subsequent writing. An important collection on the subject is Harding, ed., 1976. Important and much-discussed recent treatments of underdetermination can be found in van Fraassen 1980, ch. 2, Laudan and Leplin 1991, pp. 449–72, 1993, pp. 8–16, Kukla 1993, pp. 1–7, and Hoefer and Rosenberg 1994, pp. 592–607.
- 20 See Kiesepää 1997, pp. 21–48.
- 21 The classic work on the syntactic versus the semantic views is Suppe 1977, and his *The Semantic Conception of Theories and Scientific Realism*, which defends the semantic view, also defended in Giere 1988, and van Fraassen 1989, ch. 9. More recent work on scientific representation includes van Fraassen 2008, discussed in Ladyman et al. 2011, pp. 417–42.
- 22 Hesse 1963, and see also Frigg and Hartmann 2012, which contains many further references. On simulation see Winsberg 2009, pp. 835–45 and Section 4 below. On causal modeling see Irzik and Meyer 1987, pp. 495–514.
- 23 Niiniluoto 1998, pp. 1–29.
- 24 Hesse 1961.
- 25 Ellis 2009, Bird 2007, Chakravartty 2007.
- 26 Quine, “Natural Kinds” in his *Ontological Relativity and Other Essays*, pp. 114–38, and also in Boyd et al. 1991; Putnam, “The Meaning of Meaning” in his *Mind, Language and Reality: Philosophical Papers Vol. 2*, pp. 215–71, and “On Properties,” in Rescher et al. 1969, pp. 235–54; Mellor and Oliver 1997. Influential discussions of natural kinds include Mellor 1977, pp. 299–312, in his *Matters of Metaphysics*; Dupré 1981, pp. 66–90 and 1993, and Armstrong 1989. See also Bird 1998, ch. 3, and Bird and Tobin 2010.
- 27 Mackie 1980, and his “Causes and Conditions,” pp. 254–64 in Sosa 1975, pp. 15–38, and Sosa and Tooley 1993, pp. 33–55; Lewis 1973, pp. 556–67, and “Postscripts” and “Causal Explanation,” all in his *Philosophical Papers, Volume II*; Davidson “Causal Relations,” in his *Essays on Actions and Events*, pp. 149–62; Russell 1912, pp. 1–26, in his *Mysticism and Logic*, pp. 180–208; Cartwright 1989; Owens 1992.
- 28 Eells 1991. Lewis, “Postscripts” to “Causation,” *Philosophical Papers, Volume II*, pp. 172–213; Salmon 1984; Cartwright 1989; Mellor 1991, chs 6 and 7.
- 29 Sklar 2000; Gillett and Loewer 2001; Roberts 2008; Lange 2009.
- 30 Lewis’ discussion of laws in his *Counterfactuals* (1973) is hugely influential. See also Ramsey, “Universals of Law and of Fact,” in his *Philosophical Papers* (1978), pp. 128–32, and M. Hesse, “A Revised Regularity View of Scientific Laws,” in Mellor 1980, pp. 87–104. For a critique of regularity theories see Armstrong 1985, chs 1–5, and van Fraassen 1989, ch. 3.
- 31 Tooley 1977, pp. 667–98; Swoyer 1982, pp. 203–23; Lewis 1983, pp. 243–77; Armstrong 1985; Papineau, “Laws and Accidents,” in Wright and Macdonald 1986; van Fraassen 1989; Mellor, “Necessities and Universals in Laws of Nature,” in his *Matters of Metaphysics* (1991); Bird 1998, ch. 1; Dretske, “Laws of Nature,” in Curd and Cover 1998; Psillos 2002.
- 32 Pietroski and Rey 1995, pp. 81–110.

- 33 Nagel 1961, ch. 11; Hempel 1966, ch. 8; Fodor 1974, pp. 97–115, in his *Representations* (1981); Friedman, “Theoretical explanation,” in Healey 1981, pp. 1–16; Papineau 1993, chs 1 and 2; Dupré 1993; Cartwright 1994, pp. 279–92, in Papineau 1996; Cartwright 1999.
- 34 Magnus and Callender 2004, pp. 320–38. This paper makes a distinction between “wholesale” and “retail” approaches to the realism debate arguing for the latter.
- 35 For the history of the debates about scientific realism and a defense of it see Psillos 1996. The present discussion owes much to Psillos. See also chapters 4–8 of Ladyman 2002.
- 36 See the first chapter of Fraassen 1980.
- 37 For criticism from realists and a response from van Fraassen see Churchland and Hooker 1985. van Fraassen further elaborates his view in his *Laws and Symmetry*. See also chapter 7 of Ladyman 2002 and Monton 2007.
- 38 Harman 1965, pp. 88–95; Lipton 1991.
- 39 See his *Exceeding Our Grasp* (2006).
- 40 Lewis 2001, pp. 371–80; Lange 2002, pp. 281–5.
- 41 Hardin and Rosenberg 1982, pp. 604–15.
- 42 The most detailed and influential response to the argument from theory change is due to Psillos 1999, who combines strategies (I) and (II). For criticism see Chang 2002, pp. 902–12; Stanford 2003a, pp. 913–25, and 2003b, pp. 551–72. Elsamahi 2005, pp. 1350–60; and Lyons 2006, pp. 537–60. Other responses include Kitcher 1993, who defends a model of reference according to which some tokens of theoretical terms refer and others do not.
- 43 Cartwright 1983 and Hacking 1983.
- 44 Worrall 1989, pp. 99–124. Reprinted in Papineau 1996, pp. 139–65; Ladyman 1998, pp. 409–24. For an introduction to structural realism with extensive references see Ladyman 2009.
- 45 Brush 2007, pp. 256–9, 2008; Harker 2008, pp. 429–53
- 46 The classic discussion of explanation in philosophy of science is Hempel’s *Aspects of Scientific Explanation* (1965). Critical discussion of it can be found in the essays in Ruben 1993, and in his *Explaining Explanation* (1990). A more recent collection is Knowles 1990. B. C. van Fraassen’s much anthologised discussion of explanation in which he offers his own pragmatic account is in chapter 5 of his *The Scientific Image* (1980); and see also his *Laws and Symmetry* (1989). Widely discussed accounts of explanation in terms of unification are due to Philip Kitcher and Michael Friedman (Friedman, “Theoretical Explanation,” in Healey 1981, pp. 1–16; Kitcher 1981, pp. 507–31, in Boyd et al. 1991; Kitcher 1989). See also Salmon 1984, and “Scientific Explanation” in Salmon et al. 1999, pp. 7–41; Bird 1998, ch. 2; the essays in Sklar 2000; and Ladyman 2002, ch. 6.
- 47 <http://plato.stanford.edu/entries/mathphil-indis/>
- 48 <http://plato.stanford.edu/entries/neuroscience/>
- 49 Batterman 2013.
- 50 <http://plato.stanford.edu/entries/biology-philosophy/>
- 51 <http://plato.stanford.edu/entries/game-evolutionary/>
- 52 Parker 2011.
- 53 <http://plato.stanford.edu/entries/chemistry/>
- 54 Cleland 2011, pp. 551–82.
- 55 Ladyman et al. 2013, pp. 1–35.
- 56 A brief introductory essay and readings on Science and Medicine, Forensic Science, and Science, Race and Gender can be found in Bird and Ladyman 2012.

Bibliography

- Armstrong, D., 1989. *Universals: An Opinionated Introduction*. Boulder: Westview Press.
- , 1985. *What Is a Law of Nature?* Cambridge: CUP, chs 1–5.
- Balashov, Y. and Rosenberg, A., 2002. *Philosophy of Science: Contemporary Readings*. London: Routledge.
- Barnes, E. C., 2008. *The Paradox of Predictivism*. Cambridge: CUP.
- Batterman, R., ed., 2013. *The Oxford Handbook of Philosophy of Physics*. Oxford: OUP.
- Bird, A., 2007. *Nature's Metaphysics: Laws and Properties*. Oxford: OUP.
- , 2000. *Thomas Kuhn*. Princeton and London: Princeton University Press and Acumen Press.
- , 1998. *Philosophy of Science*. London: Routledge.
- Bird, A. and Tobin, E., 2010. "Natural Kinds." *Stanford Encyclopedia of Philosophy* (Summer edn), E. N. Zalta, ed., URL = <http://plato.stanford.edu/archives/sum2010/entries/natural-kinds>
- Bovens, L. and Hartmann, S., 2003. *Bayesian Epistemology*. Oxford: OUP.
- Boyd, R. et al., eds., 1991. *Philosophy of Science*. Cambridge: MIT Press.
- Brown, H., 1993. *Perception, Theory and Commitment*. Chicago: University of Chicago Press.
- Brush, S. G., 2007. "Predictivism and the Periodic Table." *Studies in History and Philosophy of Science Part A*, 38, pp. 256–9.
- Cartwright, N., 1999. *The Dappled World*. Cambridge: CUP.
- , 1994. "Fundamentalism vs. the Patchwork of Laws." *Proceedings of the Aristotelian Society*, 94, pp. 279–92.
- , 1989. *Nature's Capacities and their Measurement*. Oxford: OUP.
- Chakravartty, A., 2007. *A Metaphysics for Scientific Realism: Knowing the Unobservable*. Cambridge: CUP.
- Churchland, P. M., 1990. *A Neurocomputational Perspective*. Cambridge: MIT Press, ch. 12.
- , 1979. *Scientific Realism and the Plasticity of Mind*. Cambridge: CUP.
- Churchland, P. M. and Hooker, C. A., eds, 1985. *Images of Science: Essays on Realism and Empiricism*. Chicago: University of Chicago Press.
- Cleland, C. E., 2011. "Prediction and Explanation in Historical Natural Science." *British Journal for the Philosophy of Science*, 62, pp. 551–82.
- Curd, M. and Cover, J. A., eds, 1998. *Philosophy of Science*. New York: W. W. Norton.
- Davidson, D., 2001. *Essays on Actions and Events*. Oxford: OUP, pp. 149–62.
- , 1967. "Causal Relations." *The Journal of Philosophy*, 64 (21), pp. 691–703.
- Douven, I. and Meijs, W., 2007. "Measuring coherence." *Synthese*, 156, pp. 405–25.
- Dretske, F. I., 1977. "Laws of Nature." *Philosophy of Science*, 44 (2), pp. 248–68.
- Duhem, P., 1954. *The Aim and Structure of Physical Theory*. Translated by P. Wiener. Princeton: Princeton University Press, ch. 6.
- Dupré, J., 1993. *The Disorder of Things: The Metaphysical Foundations of the Disunity of Science*. Cambridge: Harvard University Press.
- , 1981. "Natural Kinds and Biological Taxa." *Philosophical Review*, 90, pp. 66–90.
- Earman, J., 1992. *Bayes or Bust*. Cambridge: MIT Press.

- Earman, J. and Salmon, M. H., 1992. "Confirmation of Scientific Hypotheses." In Salmon et al., eds, *Introduction to the Philosophy of Science*. Indianapolis: Hackett Publishing, pp. 42–103.
- Eells, E., 1991. *Probabilistic Causality*. Cambridge: CUP.
- Ellis, B., 2009. *The Metaphysics of Scientific Realism*. Durham: Acumen Publishing.
- Elsamahi, M., 2005. "A Critique of Localised Realism." *Philosophy of Science*, 72, pp. 1350–60.
- Fitelson, B., 2006. "The Paradox of Confirmation." *Philosophy Compass*, 1 (1), pp. 95–113.
- Fodor, J., 1991. "The Dogma that didn't Bark." *Mind*, 100 (2), pp. 201–20.
- , 1990. *A Theory of Content and Other Essays*. Cambridge: MIT Press.
- , 1984. "Observation Reconsidered." *Philosophy of Science*, 51 (1), pp. 23–43.
- , 1981. *Representations*. Cambridge: MIT Press.
- , 1974. "Special Sciences." *Synthese*, 28 (2), pp. 97–115.
- Fraassen, B. van, 1989. *Laws and Symmetry*. Oxford: OUP.
- Friedman, M., 1981. "Theoretical explanation." In Healey, ed., *Reduction, Time and Reality*, Cambridge: CUP, pp. 1–16.
- Frigg, R. and Hartmann, S., 2012. "Models in Science." *Stanford Encyclopedia of Philosophy* (Fall edn), Edward N. Zalta, ed. Available at: <http://plato.stanford.edu/archives/fall2012/entries/models-science/>
- Giere, R., 1988. *Explaining Science*. Chicago: University of Chicago Press.
- Gillett, C. and Loewer, B., eds, 2001. *Physicalism and Its Discontents*. Cambridge: CUP.
- Glymour, C., 1980. "Why I am not a Bayesian." *Theory and Evidence*. Princeton: Princeton University Press, pp. 63–93.
- Glymour, C. and Eberhardt, F., 2011. "Hans Reichenbach." *Stanford Encyclopedia of Philosophy* (Summer edn), E. N. Zalta, ed. Available at: <http://plato.stanford.edu/archives/sum2011/entries/reichenbach/>
- Glymour, C. N., 1980. *Theory and Evidence*. Princeton: Princeton University Press.
- Goodman, N., 1955. *Fact, Fiction, and Forecast*. Cambridge: Harvard University Press.
- Gower, B., 1997. *Scientific Method: An Historical and Philosophical Introduction*. London: Routledge, ch. 3.
- Hacking, I., 1990. *The Taming of Chance*. Cambridge: CUP.
- , 1983. *Representing and Intervening, Introductory Topics in the Philosophy of Natural Science*. Cambridge: CUP.
- , 1975. *The Emergence of Probability*. Cambridge: CUP.
- Hájek, A., 2012. "Interpretations of Probability." *Stanford Encyclopedia of Philosophy* (Summer edn), E. N. Zalta, ed. Available at: <http://plato.stanford.edu/archives/sum2012/entries/probability-interpret/>
- Hanson, N. R., 1958. *Patterns of Discovery*. Cambridge: CUP, ch. 1.
- Hardin, C. L. and Rosenberg, A., 1982. "In Defence of Convergent Realism." *Philosophy of Science*, 49, pp. 604–15.
- Harding, S. G., ed., 1976. *Can Theories Be Refuted?* New York: Springer.
- Harding, S. G. and Hintikka, M. B., eds, 2003. *Discovering Reality: Feminist Perspectives on Epistemology, Metaphysics, Methodology, and Philosophy of Science*. New York: Springer.

- Harker, D., 2008. "On the Predilections for Predictions." *British Journal for the Philosophy of Science*, 59, pp. 429–53.
- Harman, G., 1965. "Inference to the Best Explanation." *Philosophical Review*, 74, pp. 88–95.
- Hasok, C., 2002. "Preservative Realism and Its Discontents: Revisiting Caloric." *Philosophy of Science*, 70, pp. 902–12.
- Hempel, C. G., 1966. *Philosophy of Natural Science*. Upper Saddle River: Prentice Hall, ch. 8.
- , 1965. *Aspects of Scientific Explanation*. New York: Free Press, pp. 3–46.
- , 1945. "Studies in the logic of confirmation." *Mind*, 54 (213), pp.1–26.
- Hesse, M., 1980. "A Revised Regularity View of Scientific Laws." In D. H. Mellor, ed., *Science, Belief and Behaviour*. Cambridge: CUP, pp. 87–104.
- , 1963. *Models and Analogies in Science*. London: Sheed and Ward.
- , 1961. *Forces and Fields*. New York: Philosophical Library.
- Hoefer, C. and Rosenberg, A., 1994. "Empirical Equivalence, Underdetermination, and Systems of the World." *Philosophy of Science*, 61 (4), pp. 592–607.
- Horwich, P., 1982. *Probability and Evidence*. Cambridge: CUP.
- Howson, C. and Urbach, P., 1989. *Scientific Reasoning: The Bayesian Approach*. Chicago: Open Court.
- Irzik, G. and Meyer, E., 1987. "Causal Modeling: New Directions for Statistical Explanation." *Philosophy of Science*, 54, pp. 495–514.
- Kieseppä, I. A., 1997. "Akaike Information Criterion, Curve-fitting, and the Philosophical Problem of Simplicity." *British Journal for the Philosophy of Science*, 48, pp. 21–48.
- Kitcher, P., 1993. *The Advancement of Science*. Oxford: OUP.
- , 1989. "Explanatory Unification and the Causal Structure of the World." In P. Kitcher and W. C. Salmon, eds, *Scientific Explanation*. Minneapolis: University of Minnesota Press, pp. 410–505.
- , 1981. "Explanatory Unification." *Philosophy of Science*, 48, pp. 507–31, in R. Boyd et al., eds, 1991. *Philosophy of Science*. Cambridge: MIT Press.
- Knowles, D., ed., 1990. *Explanation and its Limits*. Cambridge: CUP.
- Kuhn, T. S., 1977. *The Essential Tension: Selected Studies in Scientific Tradition and Change*. Chicago and London: University of Chicago Press.
- , 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press (2nd edn, 1970).
- Kukla, A., 1993. "Laudan, Leplin, Empirical Equivalence, and Underdetermination." *Analysis*, 53 (1), pp. 1–7.
- Ladyman, J. et al., 2013. *Philosophy of Pseudoscience: Reconsidering the Demarcation Problem*. Edited by Massimo Pigliucci and Maarten Boudry. Chicago: University of Chicago Press.
- , 2009. "Structural Realism." *Stanford Encyclopedia of Philosophy* (Summer edn), E. N. Zalta, ed. Available at: <http://plato.stanford.edu/archives/sum2009/entries/structural-realism/>
- , 2002. *Understanding Philosophy of Science*, London: Routledge.
- , 1998. "What is Structural Realism?" *Studies in History and Philosophy of Science*, 29, pp. 409–24.

- Ladyman, J. and Bird, A., eds, 2012. *Arguing About Science*. London: Routledge, section 2.
- Ladyman, J., Lambert, J., and Wiesner, K., 2013. "What Is a Complex System?" *European Journal for Philosophy of Science*, 3, pp. 1–35.
- Lakatos, I., 1970. "The Methodology of Scientific Research Programmes." In *Criticism and the Growth of Knowledge*. Cambridge: CUP.
- Lange, M., 2009. *Laws and Lawmakers*. Oxford: OUP.
- , 2002. "Baseball, Pessimistic Inductions and the Turnover Fallacy." *Analysis*, 62, pp. 281–5.
- Larry, L., 1977. *Progress and its Problems*. Berkeley, CA: University of California Press.
- Laudan, L. and Leplin, J., 1993. "Determination Underdetermined." *Analysis*, 53 (1), pp. 8–16.
- , 1991. "Empirical Equivalence and Underdetermination." *Journal of Philosophy*, 88 (9), pp. 449–72.
- Lewis, P., 2001. "Why the Pessimistic Induction Is a Fallacy." *Synthese*, 129, pp. 371–80.
- , 1973. *Counterfactuals*. Oxford: Basil Blackwell.
- Lewis, D., 1986. "Postscript to Causation" and "Causal Explanation." In *Philosophical Papers, Volume II*, Oxford: OUP, pp. 172–240.
- , 1983. "New Work for a Theory of Universals." *Australasian Journal of Philosophy*, 61, pp. 243–77.
- , 1973. "Causation." *Journal of Philosophy*, 70, pp. 556–67.
- Lipton, P., 2006. *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*. Oxford: OUP.
- , 1991. *Inference to the Best Explanation*. London: Routledge.
- Lyons T., 2006. "Scientific Realism and the stratagema de divide et impera." *British Journal for the Philosophy of Science*, 57, pp. 537–60.
- Mackie, J., 1980. *The Cement of the Universe*. Oxford: OUP.
- , 1965. "Causes and Conditions." *American Philosophical Quarterly*, 2 (4), pp. 254–64, in E. Sosa, ed., 1975. *Causation and Conditionals*. Oxford: OUP, pp. 15–38.
- Magnus, P. D. and Callender, C., 2004. "Realist Ennui and the Base Rate Fallacy." *Philosophy of Science*, 71, pp. 320–38.
- Mayo, D., 1996. *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.
- Meacham, C. J. G., 2010. "Two Mistakes Regarding the Principal Principle." *British Journal for the Philosophy of Science*, 61, pp. 407–31.
- Mellor, D. H., 1991b. "Necessities and Universals in Natural Laws." In *Matters of Metaphysics*. Cambridge: CUP, pp. 136–53.
- , 1991a. *Matters of Metaphysics*. Cambridge: CUP.
- , 1977. "Natural Kinds." *British Journal for the Philosophy of Science*, 28, pp. 299–312.
- Mellor, D. H. and Oliver, A., eds, 1997. *Properties*. Oxford: OUP.
- Monton, B., ed., 2007. *Van Fraassen's Philosophy of Science* (Mind Occasional Series). Oxford: OUP.
- Musgrave, A., 1974. "Logical versus Historical Theories of Confirmation." *British Journal for the Philosophy of Science*, 25, pp. 1–23.
- Nagel, E., 1961. *The Structure of Science*. San Diego: Harcourt, Brace & World, ch. 11.

- Niiniluoto, I., 1998. "Verisimilitude: The Third Period." *British Journal for the Philosophy of Science*, 49, pp. 1–29.
- Owens, D., 1992. *Causes and Coincidences*. Cambridge: CUP.
- Papineau, D., 1993. *Philosophical Naturalism*. Oxford: Blackwell, chs 1 and 2.
- , 1986. "Laws and Accidents." In C. Wright and G. Macdonald, eds, *Fact, Science and Morality*. Oxford: Blackwell, pp. 189–218.
- Papineau, D., ed., 1996. *The Philosophy of Science*. Oxford: OUP.
- Parker, W., 2011. "When Climate Models Agree: The Significance of Robust Model Predictions." *Philosophy of Science*, 78, pp. 579–600.
- Pietroski, P. and Rey, G., 1995. "When Other Things aren't Equal: Saving Ceteris Paribus Laws from Vacuity." *British Journal for the Philosophy of Science*, 46, pp. 81–110.
- Pigliucci, M. and Boudry, M., ed., 2013. *It's Philosophy of Pseudoscience: Reconsidering the Demarcation Problem*. Chicago: University of Chicago Press.
- Pinnick, C. L., 1994. "Feminist Epistemology: Implications for Philosophy of Science." *Philosophy of Science*, 61, pp. 646–57.
- Popper, K., 1963. *Conjectures and Refutations*. London: Routledge.
- , 1959. "The Propensity Interpretation of Probability." *British Journal for the Philosophy of Science*, 10, pp. 25–42.
- Psillos, S., 2002. *Causation and Explanation*. Durham: Acumen, section II.
- , 1996. *Scientific Realism: How Science Tracks the Truth*. London: Routledge.
- Putnam, H., 1975. "The Meaning of Meaning." In *Mind, Language and Reality: Philosophical Papers Vol 2*. Cambridge: CUP, pp. 215–71.
- , 1969. "On Properties." In N. Rescher et al., eds, *Essays in Honour of Carl Hempel*. New York: Springer, pp. 235–54.
- Quine, W. V., 1969. "Natural Kinds." In *Ontological Relativity and Other Essays*. New York: Columbia University Press, pp. 114–38.
- , 1953. "Two Dogmas of Empiricism." In *From a Logical Point of View*. Cambridge: Harvard University Press, pp. 20–46.
- Ramsey, F., 1978, "Universals of Law and of Fact." In D. H. Mellor, ed., *Philosophical Papers*. Cambridge: CUP, pp. 128–32.
- Roberts, J., 2008. *The Law-Governed Universe*. Oxford: OUP.
- Ruben, D., 1990. *Explaining Explanation*. London: Routledge.
- Ruben, D., ed., 1993. *Explanation*. Oxford: OUP.
- Russell, B., 1919. *Mysticism and Logic*. London: Longmans, pp. 180–208.
- , 1912. "On the Notion of Cause." *Proceedings of the Aristotelian Society*, New Series (13), pp. 1–26.
- Salmon, W. C., 1999. "Scientific Explanation." In W. C. Salmon et al., eds, *Introduction to the Philosophy of Science*. Indianapolis: Hackett Publishing, pp. 7–41.
- , 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.
- Shapere, D., 1982. "The Concept of Observation in Science and Philosophy." *Philosophy of Science*, 49 (4), pp. 485–525.
- Sklar, L., ed., 2000. *Explanation, Law and Cause*. London: Routledge.
- Sosa, E. and Tooley, M., eds, 1993. *Causation*. Oxford: OUP, pp. 33–55.
- Stanford, P. K., 2003b. "Pyrrhic victories for scientific realism." *Journal of Philosophy*, 11, pp. 551–72.

- , 2003a. "No Refuge for Realism: Selective Confirmation and the History of Science." *Philosophy of Science*, 70, pp. 913–25.
- Suppe, F., 1989. *The Semantic Conception of Theories and Scientific Realism*. Champagne: University of Illinois Press.
- , 1977. *The Structure of Scientific Theories*. Champagne: University of Illinois Press.
- Swinburne, R., 1971. "The Paradoxes of Confirmation—A Survey." *American Philosophical Quarterly*, 8 (4), pp. 318–30.
- Swoyer, C., 1982. "The Nature of Natural Laws." *Australasian Journal of Philosophy*, 60 (3), pp. 203–23.
- Tooley, M., 1977. "The Nature of Laws." *Canadian Journal of Philosophy*, 7, pp. 667–98.
- Urbach, P., 1987. *Francis Bacon's Philosophy of Science: An Account and a Reappraisal*. Chicago: Open Court.
- van Fraassen, B. C., 2008. *Scientific Representation: Paradoxes of Perspective*, Oxford: OUP. Discussed in Ladyman, J., Bueno, O., Suárez, M., and van Fraassen, B. C., 2011. "Scientific Representation: A Long Journey from Pragmatics to Pragmatics." *Metascience*, 20, pp. 417–42.
- , 1989. *Laws and Symmetry*, Oxford: OUP, ch. 9.
- , 1980. *The Scientific Image*, Oxford: OUP, ch. 2.
- Winsberg, E., 2009. "Computer Simulation and the Philosophy of Science." *Philosophy Compass*, 4, pp. 835–45.
- Woolhouse, R., 1988. *The Empiricists*, Oxford: OUP, ch. 2.
- Worrall, J., 1989. "Structural Realism: The Best of Both Worlds?" *Dialectica*, 43, pp. 99–124. Reprinted in D. Papineau, ed., 1996. *The Philosophy of Science*. Oxford: OUP, pp. 139–65.
- Zahar, E., 1989. *Einstein's Revolution: A Study in Heuristic*, Michigan: University of Michigan Press.
- , 1973. "Why did Einstein's Programme Supersede Lorentz's?" *British Journal for the Philosophy of Science*, 24 (3), pp. 95–123, 223–62.

20 Philosophy of Physics

Barry Loewer

1 What Are Physics and the Philosophy of Physics?

The philosophy of physics concerns the philosophical foundations of specific theories in physics—classical mechanics, electrodynamics, statistical mechanics, quantum mechanics, relativity—and also more general philosophical issues concerning the nature and aims of physics, the metaphysical nature of fundamental laws, matter, fields, space and time, the direction of time, objective chance, reduction, and causation. Philosophy of physics is not only a major field in its own right but is also important for other parts of philosophy, especially epistemology and metaphysics. It is significant for epistemology because physics is the source of our most general and fundamental knowledge of the natural world and in particular knowledge of those parts of the world that are not accessible to observation. It is essential to metaphysics since it provides the best accounts we have of the universe's fundamental ontology and explanatory structure. Physics is arguably “the royal road to metaphysics.”¹ This chapter discusses a few of the most interesting and accessible issues in the philosophy of physics and also some points at which developments in the philosophy of physics are relevant to epistemology and metaphysics. However, the reader is forewarned that the subject of philosophy of physics is vast and a great deal of it is technical. Perforce my selection of topics and emphasis are idiosyncratically shaped both by my particular interests and a desire to keep technicalities to a minimum.

Greek natural philosophers (Leucippus, Democritus, Epicurus, and Aristotle) laid down many of the ideas that are central to the development of physics. Specifically, they set the primary subject matter of physics as regularities in the *motions* of material objects. They also introduced the idea that some motions (e.g. circular, straight, etc.) are *natural*, which turned out to be very influential much later. Greek and Roman atomists held that material objects and processes are composed of a few kinds of fundamental unobservable parts (atoms) and that the behavior of complex phenomena is to be accounted for in terms of the arrangements and behavior of their parts. The scope of physics has been greatly extended over the centuries but the unobservable ontological posits

that have been introduced (e.g. fields, space-times, wave functions, elementary particles, dark matter, etc.) ultimately earn their keep by the roles they play in accounting for the motions of quotidian material objects.

Although ancient Greek and Egyptian philosophers and astronomers described some motions with great accuracy (e.g. Ptolemy's description of the motions of the planets) it was not until the seventeenth century that the mathematics needed to describe motion and changes in motion (primarily the calculus) was sufficiently developed and employed. Around the same time the modern conception of "law of nature" as expressed by differential equations emerged and it was recognized that the motions of celestial and terrestrial objects were governed by the same laws. These developments were essential elements to the creation of physics as an exact mathematical science whose aim is to explain all motion in terms of a few simple laws and a few ultimate constituents of matter. The crowning achievement was Isaac Newton's formulation in his *Principia Mathematica* of a few mathematical laws that he thought govern the motions of all material objects and his use of these principles to explain and thus unify terrestrial and celestial motions.

2 Newtonian Mechanics

If we understand Newtonian mechanics as an attempt to formulate a fundamental and complete theory of the world it claims that the universe consists of a fixed number of indestructible point particles that reside in a 3-D Euclidian space and move relative to the flow of time. Absolute space and time as they are called is the stage on which the particles perform. Particles have intrinsic properties (mass, electric charge, and so on) and extrinsic relations derived from their locations in space and time. Thus two particles are a mile apart at time t if they are located at points of space that are a mile apart and a particle is moving at an average speed of one mile/hour between times t and t' if the distance the particle traverses through absolute space divided by the duration from t to t' is one mile/hr.² The velocity at an instant is a vector quantity equal to the derivative of distance (in a direction) with respect to time and acceleration is the second derivative with respect to time.³ The guiding idea of Newtonian mechanics is that a particle's *natural* motion (motion when no force is acting on it) is to continue in the same direction in a straight line at a constant speed. A change in a particle's speed or direction (acceleration) is due to the forces acting in that direction on the particle. The notions of straight line and constant velocity, etc. are specifiable in terms of the Euclidian structure of absolute space and time and forces acting on a particle are characterized in terms of laws.

Newtonian mechanics contains three types of laws; (a) laws of force that specify the forces acting on particles depending on their intrinsic properties and spatial arrangements, (b) laws concerning the nature and compositions of forces, and (c) a dynamical law that specifies how the total force on a particle affects its motion. Forces are also vector quantities. Examples of each kind of law are (a) Newton's law of gravitation $F(1,2)=Gm_1m_2/r(1,2)^2$, (b) the total force on a particle is the vector sum of individual forces on it, and (c) $F(p)=ma(p)$ where $F(p)$ is the total force on particle p and $a(p)$ is the acceleration of p .

Newtonian mechanics is *complete*, *deterministic*, and *time reversible*. Newton did not know all the kinds of particles and forces there are but we can suppose that when they were discovered his account would be *complete* in the sense that an inventory of what particles exist, their intrinsic properties, their trajectories throughout all time, and the totality of forces and their associated laws represent absolutely everything that could be said about the physical history and operation of the universe in the fundamental language of physics. Insofar as macroscopic entities and properties (planets, mountains, clouds, butterflies, etc.) supervene on fundamental physical history and fundamental laws macroscopic history would also be completely specified.

Newtonian mechanics is two-way *deterministic* in the sense that given the classical mechanical state at any time t (i.e. the positions and velocities of each particle and their intrinsic properties at t), the entire future and past history of the universe in every detail is determined by the laws.⁴ Newtonian mechanics is time reversible in that for any sequence of states that is compatible with the laws there is a temporally reversed sequence that is also compatible with the laws. We will have more to say about time reversibility when we discuss the direction of time and statistical mechanics.

In order to apply Newtonian mechanics to macroscopic objects, for example, the moon, projectiles, springs, and so on physicists need to make certain idealizing assumptions. For example, to apply the laws to planetary motions it is assumed that although a planet is an arrangement of a very large number of point particles persisting in more or less that arrangement it can be treated as a single "point particle" located at the planet's center of mass whose mass is the sum of the masses of all the particles that compose the planet, and that the only relevant forces are gravitational originating in the sun and other objects in the solar system. On the basis of such idealizations one can derive Kepler's laws, Galileo's laws of free fall, the law of the pendulum, Hooke's law, and so on. The explanatory unification brought about by Newtonian mechanics is nothing short of mind boggling!

On its face the fundamental ontology of Newtonian mechanics consists of space, time, material point particles (and their intrinsic properties/quantities),

and perhaps also forces and laws.⁵ If Newtonian theory is understood merely as an *instrument* for making predictions (e.g. of the apparent motions of the planets in the night sky) then one need not be too concerned about understanding this ontology or how it is that the macroscopic world is constituted by or supervenes on the lawful motions of material point particles. But if one understands Newtonian mechanics in *realist* fashion as an attempt to specify fundamental reality then questions concerning the nature of this ontology arise that are quite difficult. A good deal of the history of physics from the seventeenth century to the present can be seen as a battle between (sometimes more subtle forms of) instrumentalist and realist understandings of the aims and proposals of physics. The issue of whether physical theories can be understood in realist fashion as providing an account of the world will come up a number of times in the ensuing discussion.

In Newtonian mechanics point particles constitute the substance of the physical world. Mass (resistance to acceleration) is had by all particles and is what gives the world its heft. Presumably, there are particles that are identical in the values of their intrinsic properties and so differ only in their locations in their trajectories in space. Thus Newtonian mechanics violates some versions of the principle of the identity of indiscernible.⁶ A Newtonian would say “so much the worse for the principle.” An ordinary physical object, for example, a stone, consists of many particles (perhaps with different masses and other intrinsic properties) arranged in a particular configuration and held together by forces of some kind. How the whole macroscopic world and especially so-called secondary qualities and mental phenomena are constituted by or supervene on or emerge from configurations of point particles is, to say the least, extremely difficult and puzzling. That discussion is beyond the scope of this article although, as we will see when we come to quantum mechanics, the relation between mental and physical has a way of intruding itself in to the philosophy of physics.

Newton thought of absolute space as an infinite 3-D Euclidian manifold completely uniform in its nature. Its points or locations persist over time so that it is an objective matter of fact whether a particle occupies the same or different locations at different times (i.e. whether the particle has changed location). Absolute space is a “substance” insofar as it exists independently whatever material particles inhabit it. On the Newtonian conception it would be possible for there to exist space without any particles although particles by their nature must have location in space. Newton held that time “of itself, and from its own nature, flows equably without relation to anything external.”⁷ Whatever “flow” may mean it is clear that he held that time exists independently of the material contents of space and that its “flow” determines the temporal ordering and the durations between all events (e.g. the time between two collisions). While he thought of both space and time as “absolute” he also

thought of them as fundamentally different since time but not space “flows” and this flow provides time with an intrinsic direction.

Newton’s views of space and time strike some as commonsensical. However, soon after he formulated his mechanics a controversy arose over whether the existence of absolute space and time are really required in order to account for motion and even whether they are comprehensible. Leibniz argued that absolute space offended theology and the principle of sufficient reason since God would have no reason to place the particles of the universe in one part of space rather than another. It also offends Leibniz’s principle of the identity of indiscernibles since distinct points in absolute space are qualitatively indistinguishable. These controversial metaphysical principles aside there are reasons to worry about absolute space that are more closely connected to Newtonian mechanics itself. In absolute space and time there are matters of fact concerning the spatial distance between events when they occur at different times and thus a matter of fact about a particle’s absolute velocity. But Newtonian theory implies that it is impossible to empirically determine absolute position and absolute velocity.

Newtonian space+time determines a frame of reference that can be characterized in terms of a coordinate system. By selecting an arbitrary point in absolute space and an arbitrary point in time as the origin, three mutually perpendicular directions in space and units of distance and duration each particle is provided with an address at each time. We can also define frames of reference (and coordinates) whose origin is moving uniformly with respect to the absolute frame. These are called “inertial frames.”⁸ Both Newton and his critics realized that the Newtonian laws are invariant with respect to inertial frames in that they predict exactly the same forces, relative velocities, and accelerations in every inertial frame. It follows that there are no measurements or observations we can make that will tell us whether any particular inertial frame corresponds to the absolute frame and so there is no way to determine a particle’s absolute velocity or the distance between two nonsimultaneous events. The best we can do is measure the distances between particles, their relative velocities at a time, and changes in their relative velocities. It seems that by positing absolute space and time Newton allows for “facts” that are irremediably inaccessible. One does not have to be much of an empiricist to find such “facts” suspect. Further reflection could lead one to think that absolute space is a spooky kind of entity whose only job may seem to be to provide relative distances and velocities for material objects.

Worries along these lines motivated relationist reworkings of Newtonian mechanics. Relationists about space reject the existence of absolute space and replace it with fundamental distance relations between particles. On a relationist account, instead of specifying the positions of particles relative to absolute space+time a complete inventory need only specify the *relative* distances

between particles at each time. The relationist reworking of Newtonian mechanics is an attempt to specify the laws making use only of spatial relations that has the consequence of specifying exactly those spatial relations that are compatible with Newton's laws. To get a feel for the dispute will briefly describe (in updated form) one highlight in the early history of the debate between absolute and relationist conceptions of space.

Newton argued that even though it is not possible to physically measure the absolute positions and velocities of particles absolute space is nonetheless required for the proper formulation of the mechanical laws and in particular to account for accelerations. Here is a thought experiment based on Newton's famous bucket argument and his related discussion of spinning globes that is intended to show this.⁹ Consider a universe that consists entirely of two balls attached to opposite ends of a spring whose relaxed length is k . Imagine also that there is a time $t(0)$ at which the length of the spring is greater than k and that no two particles (in the balls or the spring) are changing with respect to their mutual distances. On Newton's account the future motion of the balls and spring, whether it will oscillate or remain stretched, depends on whether the spring is rotating with respect to absolute space. If the spring is stationary, it will oscillate, but if it is rotating at just the right speed, it will remain stretched. The problem for relationism is that there is no fact at $t(0)$ admitted by relationism concerning whether the spring is rotating or stretched since all the mutual distances among the particles and all the rates of change of distances are the same in either case. It follows that the state of the spring and balls at $t(0)$ is not sufficient to determine its state at subsequent times. Thus relationism cannot distinguish two different possibilities at $t(0)$ that lead to different subsequent behavior and consequently the relationist version of Newtonian mechanics is not deterministic.¹⁰

Relationists can respond by granting the failure of determinism in this simple universe but claim that this is of no consequence since the actual universe contains many more entities and relative to them there is a fact at each time as to whether a spring in the above condition is rotating. Ernest Mach proposed that acceleration should be defined relative to the "the fixed" stars or the bulk matter in the universe. In effect his suggestion to replace the frame that is determined by absolute space with the frame relative to which the fixed stars are not rotating. It follows that while on Newton's conception there may be a fact as to whether the universe as a whole is rotating in absolute space Mach's account stipulates that the universe (the fixed stars) is not rotating.¹¹ So it not only rejects "unwanted" possibilities but also possibilities that seem genuine. A further, more peculiar, consequence of Mach's relationism involves a kind of nonlocality since whether or not the spring will oscillate depends on the distribution of the fixed stars.

There are other relationist responses to Newton's argument for absolute space.¹² All have costs since by eschewing absolute space they introduce complications in the formulation of the laws. Whether one is willing to pay the costs depends on how strongly moved one is by the empiricist considerations offended by absolute space and how strongly one prizes simplicity of laws. In any case, the whole issue of relationism versus absolutism looks very different from the perspective of contemporary physics, as we will see.

Absolute time in Newtonian mechanics is even more mysterious than absolute space. Newtonian absolute time would continue to flow even in a completely empty universe. Also there are possible worlds that match each other in the temporal sequences of all events but which systematically differ in the time intervals (as determined by the flow of absolute time) between events. The empiricist worry is that there is no way to determine which universe is actual and so one may wonder whether these are real possibilities. Further, it is not obvious how one can justify the claim that those physical mechanisms we consider to be good clocks actually measure duration of Newtonian absolute time or that we should care whether or not they do. Rather, good clocks measure relative durations (e.g. how many times the clock ticks while the earth revolves around on its axis). If clocks are coordinated with each other whether or not they are coordinated with the flow of absolute time seems irrelevant. Such thoughts have led to attempts to formulate mechanics without absolute time but only in terms of temporal duration between actual (or possible) events or even just temporal orderings of events.¹³ Again there is a question whether this program can succeed and if it does whether the complexities are worth the cost.

3 Statistical Mechanics and Time's Arrows

Newton attempts to capture the idea that time has a direction by way of the metaphor that "time flows." It is not at all clear what he meant by this metaphor but it does suggest certain metaphysical views about time.¹⁴ The metaphysics of time is controversial but it is not controversial that our world is full of temporally asymmetric processes, so-called arrows of time; for example, the melting of ice, the growth of plants, the life of the stars, causes precede effects, that we have records of the past but not the future, that we can influence the future but never the past, and so on. These asymmetries are pervasive and seem lawful but there is a problem accounting for them in Newtonian mechanics since its laws are oblivious to temporal asymmetries. For every sequence of states (positions and velocities of particles) that is compatible with the laws there is a temporally reversed sequence (obtained by reversing

the direction of velocities) that is also compatible with the laws. It follows that for every possible trajectory of positions of particles there is a temporally reversed sequence of positions. So if Newtonian mechanics were a correct account of our world then it would also be a correct account of a world in which temporal processes are reversed, for example, ice cubes grow bigger, people grow younger, and so on. It appears that the Newtonian laws cannot themselves explain the lawful temporally asymmetric processes. Although Newton's talk of time flowing introduces a directionality it is unclear what this has to do with the pervasiveness of temporally asymmetric processes. As far as Newtonian mechanics is concerned time could flow in the same direction whether the sequence of events is the melting of an ice cube or the spontaneous growth of an ice cube out of warm water.

The problem of reconciling temporally asymmetric processes with Newtonian mechanics became especially urgent to physics during the nineteenth century as physicists took seriously the idea that matter is composed of atoms and so that Newtonian Mechanics or something very much like it really could be the fundamental theory of the world. At the same time a science of macroscopic phenomena (involving systems composed of many atoms), thermodynamics, was developing that has smack right in the middle of it a temporally asymmetric law—the so-called second law. The second law, as it was first formulated, says that the entropy of an energetically isolated macroscopic system never decreases and typically increases over time until the system reaches thermodynamic equilibrium. Entropy and equilibrium are thermodynamic properties of macro systems that are characterized in terms of their relationships with other thermodynamic quantities; energy, pressure, work, temperature, etc. Roughly, the entropy of a system is inversely related to the quantity of useful work in the system and a system at equilibrium is one in which there is no useful work to be gotten out of the energy in the system. For example, the process in which a hot gas in a piston chamber is allowed to expand by pushing the piston is one in which work is extracted as the piston moves (the work can drive the wheels of a car) and entropy of the entire system increases as the gas expands and cools.

The problem of squaring thermodynamics with Newtonian mechanics is that since the latter deterministically accounts for the motions of all particles in terms of a temporally symmetric law there seems to be no room (without violating those laws) for an additional temporally asymmetric dynamical law governing particles and their motions. The first big steps toward reconciling the second law with Newtonian mechanics were taken by Ludwig Boltzmann. The upshot of his years of investigation is this: Boltzmann characterized the thermodynamic properties of a macro system, pressure, temperature, energy, entropy, equilibrium, etc. in terms of classical mechanical quantities (position, momentum, total energy, etc.) and a

measure (the standard Lebesgue measure) over the set of possible states.¹⁵ He then observed that even though there are infinitely many entropy-decreasing (toward the future) micro states that realize a nonequilibrium system (e.g. an ice cube in warm water) that evolve into lower entropy states (ice cube grows bigger) such states are, in a certain sense, *rare*. The sense is that on the standard measure the measure of the set of micro-states realizing the thermodynamic condition of an isolated nonequilibrium system that is entropy-decreasing is very small. Further, the measure of the set of entropy-decreasing states in small neighborhoods of typical micro states is also very small. His next step was to construe the measure as specifying a probabilities. It follows that the conditional probability of a system in a nonequilibrium macro condition *M* being in a micro state that lies on an entropy-increasing trajectory is approximately 1.¹⁶ So it appeared that Boltzmann explained how the temporally asymmetric second law can be reconciled with the temporally symmetric fundamental dynamical laws.

However a problem was soon noticed (by Loschmidt, Zermelo, and others) with Boltzmann's proposal. As a consequence of the temporal symmetry of the fundamental laws the uniform probability distribution applied to a system at time *t* in macro condition *M* entails that the probability that the entropy of the system was greater at times *prior* to *t* also is approximately 1. Boltzmann's probability assumption entails that very likely the ice cube in an isolated Martini glass was smaller an hour ago and even earlier was entirely melted (assuming that the martini glass has been isolated during that *t*). More generally, Boltzmann's probability posit applied to the macro state of the universe at time *t* entails that it is likely that its entropy was greater at both later and earlier times. Of course this is absurd.¹⁷ If we come upon an ice cube in a martini glass that we know has been sitting isolated in a warm room for an hour we can be certain that the ice cube did not spontaneously arise out of warm water but was previously larger. So, while on the one hand, Boltzmann's probability posit apparently accounts for entropy increasing toward the future, on the other hand, it entails the absurdity that entropy was greater in the past. This is the "reversibility paradox."

The history of statistical mechanics is littered with responses to the reversibility paradox. One response is to construe the Boltzmann probability only as advice for making predictions but refrain from using it for retrodictions. This avoids the paradox but in common with other instrumentalist proposals elsewhere in the sciences it leaves us completely in the dark as to why the prescription works.¹⁸ This is not the place for a survey of various other attempts to ground the second law while avoiding the paradox so I will simply describe a proposal developed by David Albert (though it has many precedents) since, in my view, it is the most promising.¹⁹ It turns out that this proposal has profound consequences not only for the second law but also for times' other arrows.

It is generally believed on the basis of cosmological observation and theory that the state of the universe at or right after the big bang has very low entropy.²⁰ Call the very low macro state at this time $M(0)$. Albert proposes that it is a law that there is a uniform probability distribution over the possible micro states that can realize $M(0)$.²¹ So according to Albert there are three ingredients to the fundamental theory of the world.²²

- (a) The fundamental dynamical and force laws (in our discussion so far these are the Newtonian laws);
- (b) The claim that the initial macro state is $M(0)$ and that the entropy of $M(0)$ is very tiny (Albert calls this "Past Hypothesis" (PH));
- (c) A law specifying a uniform probability over physically possible micro states.²³

These three ingredients provide a probability map of the universe since they entail a probability distribution (or rather probability density) over the set of all possible micro-histories of the universe compatible with $M(0)$. This solves the reversibility paradox and explains the second law. Here is how. It follows from (a), (b), and (c) that the probability distribution over the micro states (and histories) of a system in state $M(t)$ is conditionalized on $M(t)$ and *also* $M(0)$. The measure of the set of micro states that realize $M(t)$ on the uniform distribution that are entropy-increasing in both temporal directions from t is practically 1. But conditionalizing on the very low entropy macro state $M(0)$ excludes all but a set of tiny measure of those realizers of $M(t)$ whose entropy increases to equilibrium toward the past. It thus blocks the argument that gave us the reversibility paradox. Further, it entails that conditional on the macro state at each moment prior to the universe reaching equilibrium it is overwhelmingly likely that entropy increases in the temporal direction away from the big bang.²⁴

The probabilistic version of the second law not only says that the entropy of the whole universe likely increases (or rather is likely to never decrease) as long as the universe is not yet at equilibrium but also that this holds for typical subsystems, for example, an ice cube in a glass of warm water under a wide variety of conditions. Here is a rough "seat of the pants" argument that this obtains. Suppose that S is a small subsystem of the universe that at time t "branches off" from the rest of the universe to become more or less energetically isolated and that the macro state of S is $m(t)$. We can think of the micro state of S as being selected "at random" conditional on $m(t)$ from the macro state of the universe $M(t)$. Since "almost all" (i.e. measure almost 1) micro states realizing $M(t)$ are entropy-increasing "almost all" of those realizing $m(t)$ will also be entropy-increasing; that is, $P(\text{entropy } S \text{ increases}/m(t) \& M(0))$ is approximately 1. Of course this does not mean that it is likely that the entropy

of *every* subsystem of the universe is likely to increase. Some subsystems are interacting with other parts of the universe so as make entropy decrease likely (e.g. the glass of water in the freezer). Or a system may be specially prepared so that even when it becomes isolated its entropy will very likely decrease.²⁵ In these cases the second law will be violated. But that is as it should be. The job is to ground the second law *insofar as* the second law is correct and, arguably, Albert's account does that.

Statistical mechanics probabilities help with a problem we overlooked earlier. As we observed when Newtonian mechanics (or any successor mechanics) is applied to macroscopic phenomena, for example, the motions of the moon, physicists assume that the moon can be represented by a particle at its center of mass. But in fact there are possible arrangements of the particles that comprise the moon that are compatible with its macroscopic state and in which the moon ejects some particles at great velocity and flies out of its orbit. If the particles are so arranged the idealization is incorrect. But physicists neglect these aberrant arrangements. Statistical mechanics justifies this neglect since the probability of aberrant states is miniscule.

It is worth reflecting for a paragraph or two on some of the philosophical implications of statistical mechanics. First, it suggests the possibility that all "times arrows," are grounded in the temporal asymmetry introduced by points (a), (b), and (c). If this could be shown then we might consider dropping the metaphor of "time's flowing" since it would play no role to account for the temporal asymmetries.²⁶ An objection to the account is that it presupposes the past/future distinction rather than explain it since it says that the big bang state that occurred 13.7 billion years or so in the past has very low entropy. But this is a mistake. It specifies that there is a very low entropy macro condition $M(0)$ at the time of the big bang and no similar very low entropy condition at any other time between this event and the time the universe reaches equilibrium. The macro condition at the time of the big bang will earn its name as the "*Past Hypothesis*" if it can be shown that the other arrows of time are aligned with the entropic arrow entailed by the account. That is, if it can be shown that the account not only explains the second law but also explains the temporal asymmetries of knowledge and influence, why the past seems closed and the future open, etc. on the assumption that the temporal direction of the big bang is the past then it will provide a *scientific* account of the past/future distinction.

A second consequence, both for physics and philosophy is the introduction of probabilities into physics. Statistical mechanics posits a probability distribution over the trajectories of possible states of the universe. Exactly what does probability mean here? Since it is assumed that the dynamical laws are deterministic the answer must be compatible with determinism. We will see that an issue arises again in quantum mechanics that has both

deterministic and indeterministic interpretations. It is no understatement to say that there is no consensus within philosophy of physics concerning the metaphysics of probability whether it occurs together with deterministic or indeterministic laws.²⁷

Third, the status of the PH raises some interesting epistemological questions. How do we know that the condition of the universe 13.7 billion years ago was one of very low entropy? I mentioned that this is the conclusion of cosmologists. But if statistical mechanics is correct then on Albert's account that very condition is required to ground the justifiability of the inferences that lead to this conclusion since without it the statistical mechanical distribution entails higher entropy past. Further, some physicists and philosophers believe that the PH is in some sense very unlikely since it is such a low entropy state it cries out for explanation.²⁸ Some have argued that such an unlikely event requires an explanation from outside of physics in terms of theology. Others have proposed multiverse accounts on which our universe bubbles off from a "mother universe." Alternatively, one might look more closely at the question of whether the low entropy condition PH really "requires" an explanation in any more urgent sense than any other fundamental law.²⁹

4 Theories of Relativity

During the nineteenth century it became clear to physicists that a number of phenomena—light, magnetism, and electricity—were difficult to fit into classical mechanics. Newton thought that light consisted of particles that interacted by way of some kind of force with material particles but experiment showed that light behaves more like waves than particles. Like water waves light generates interferences and is associated with frequencies and amplitude. Waves require a medium in which to propagate so it was hypothesized that space was occupied by a kind of ethereal substance dubbed "the ether." So it seemed that ether should be added to the fundamental furniture of the world.

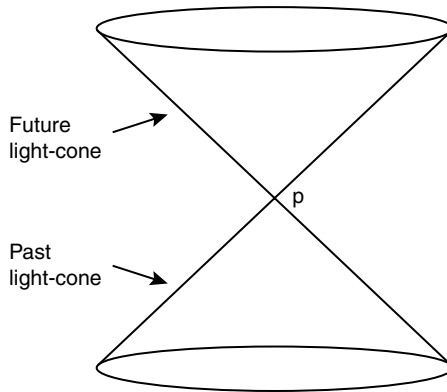
Electric and magnetic forces were found to obey laws similar to the gravitational law although the behavior of moving electrically and magnetically charged particles turned out to be complicated. There is a connection between moving charges (electricity) and light as, for example, manifested by lightning. James Maxwell proposed that electric and magnetic forces are aspects of a single electromagnetic force that itself can best be characterized in terms of a field—the electromagnetic field. Mathematically, a field is characterized by an assignment of numbers and vectors (or other mathematical objects) to points of space. At first fields were thought of merely as a convenient way of specifying the forces produced by or would be produced by a particle P on other

particles that might be located at various positions relative to P. But in the theory of electromagnetism developed by Maxwell the electromagnetic field takes on a life of its own (there are solutions in which there are no charged particles). Further it was hypothesized that light, and other kinds of radiation can be understood as waves propagating in the electromagnetic field. The field itself seemed to be either a property of the ether or perhaps a fluid-like entity that plays the role of the ether. So if a metaphysician at the end of the nineteenth century were to look to physics for the account of the fundamental ontology of the world he might conclude that it consists of 3-D space and time (the arena), particles (some charged), fields (and the ether), the fundamental dynamical laws, and the laws of statistical mechanics.³⁰

However, around the beginning of the twentieth century it became clear that there are deep problems unifying these ontologies and laws. Newtonian mechanics requires that the laws be the same in all inertial frames as characterized in terms of invariance under Galilean transformations. But Maxwell's laws explicitly entail that the speed of light is independent of the speed of the light's source and so are not Galilean invariant. Further, a famous experiment by Michelson and Morley showed that the speed of light is independent of the earth's motion thus suggesting that it is the Newtonian laws that would have to be adjusted.

To effect reconciliation between the Newtonian and Maxwellian laws Lorentz modified the class of transformations that characterize inertial frames.³¹ This modification entails that while there is a true speed of light relative to absolute space and time the measurements of the speed of light will yield the same result in all inertial frames. Lorentz's theory entails that clocks and measuring rods in motion relative to absolute space systematically "mal-function" in just such a way as to imply this result.³² As was the situation in Newtonian mechanics it is impossible to determine one's motion relative to absolute space but its existence is required to formulate the theory.

Einstein's special theory of relativity (SR) proposes a very different explanation of the constancy of the speed of light. SR rejects that there is an absolute distinction between space and time. In its stead it posits that the arena of the world is a 4-D manifold called "Minkowski space-time." The geometry of Minkowski space-time is characterized in terms of a "distance" between space-time points that specifies the paths that a beam of light can take when moving in a vacuum. Given a point p in Minkowski space-time there is a 4-D cone formed from the paths that light can take when emitted from that point and a cone corresponding to the paths that light can take to converge on that point. Paths entirely within light cone correspond to paths (world lines) that a material particle can have and straight lines within a cone correspond to inertial paths (paths taken by particles that are not accelerating).



The Newtonian and Maxwellian laws formulated in terms of this space-time entail that the speed of light is independent of the motion of its source since light emitted at a point travels on the surface of the forward light cone associated with that point irrespective of the motion of the source. This consequence is required to reconcile mechanics and electrodynamics. But there are other, astonishing consequences, of SR. For example, unlike Newtonian space in which all events are completely ordered in time, in SR only those events that lie within the past and future light cones of a point are temporally ordered. Two events that lie outside of each other's light cone are not temporally comparable (they are said to be "space-like" related). Newton's view that time flows equably throughout all of space is thus rejected. The properties mass, length, speed, and shape that have application at a time in Newtonian absolute space+time are strictly never instantiated in SR. That is, there is no such thing as *the* length of a rod, or *the* shape of an object, or *the* velocity of a body at a time *t* and for that matter there is no such thing as *the* time of an event.³³

However, length, simultaneity, etc. have surrogates that are characterized relative to a Lorentzian inertial frame. In other words, while there is no true length of a rod in Minkowski space-time there is the length of a rod relative to a Lorentz frame. A Lorentz frame is a coordinate system defined in terms of a family of inertially co-moving clocks (i.e. clocks moving on parallel paths). It is possible to "synchronize" the co-moving clocks thus specifying a time coordinate and to lay down 3-D coordinates specifying spatial locations. These then define length, duration, velocity, and so on, but only relative to the coordinate system determined by this family. Another family of co-moving clocks determines another coordinate system. Relative to one coordinate system an object that is spherical and moving at velocity *v* at time *t* (in that frame) will have a different shape and velocity relative to another frame. Lorentz frames

themselves of course have no fundamental existence. Rather they and the surrogate relative concepts are introduced as part of an explanation of how the manifest world—the macro world as it seems to us under normal conditions—appears to have a Newtonian structure when in fact the fundamental space-time is Minkowskian.

SR has many astonishing (and empirically confirmed) consequences. The most famous is the equivalence of mass and energy expressed by $E=mc^2$ but the most astonishing is the relativity of temporal duration to motion relative to a frame and to paths traveled through space-time. For example, particles moving at a very high velocity (relative to our frame of reference) have longer (relative to our frame) half-lives.³⁴ Another is that clocks that move from one point in space-time to another on different paths will record different times. If twins meet at point *p* and one speeds off to the stars at high velocity and then returns to meet his twin (so they travel on different paths in space-time) he will have aged less than the stay-at-home twin. This “paradox” is explained by the fact that the twins travelled on different paths through space-time (clocks are to “distance” traveled in Minkowski space-time as odometers are to distance traveled in space).

In Newtonian mechanics mass plays two different roles. On the one hand it measures the resistance of a body to being accelerated by force (inertial mass) and on the other it determines the extent to which any material body exerts an attractive gravitational force on other bodies (gravitational mass). These are equivalent in that the ratio of any body’s inertial mass to its gravitational mass is a constant. Because of the equivalence bodies with different masses fall in a gravitational field at the same rate (as famously demonstrated by Galileo in Pisa). In Newtonian mechanics this equivalence seems completely coincidental. Einstein devised his generalized relativity theory (GR) to explain the equivalence by building gravitation into the geometry of space-time. The basic idea is that the distribution of matter and energy determines the curvature of space-time and the curvature determines the trajectories of matter and radiation in accordance with certain laws (Einstein’s field equations). The resulting space-time geometry may be non-Euclidian (as, for example, the 2-D geometry on the surface of a sphere or the inside of a saddle). In a non-Euclidean space-time inertial paths (world lines of bodies on which there is no force) traveled by bodies are curved geodesics. Thus two nearby objects fall at the same rate in a gravitational field not because of a gravitational force (there is no such thing in GR) but because they are traveling on similarly curved (in space-time!) inertial paths. GR not only explains the equivalence of inertial and gravitational mass but it also does away with the issue of the mechanism of gravitation—how gravitational force can act over great distances—by eliminating gravitational force in favor of the geometrical structure of space-time.

There is an enormous variety of space-times compatible with GR. Among these are universes in which space-time is closed so that a light beam sent in a direction eventually returns to its origin, which contain black holes (masses so dense that no light can escape), space-times with causal loops in which it appears that time travel is possible, space-times in which there is an “initial” point in time (as the big bang is usually thought to be), space-times that expand (and contract), and so on. GR is the framework in which cosmological theories about the origin and nature of the universe are formulated.³⁵

Space and time in SR and even more so in GR are radically different from Newton’s view of space as the stage of the universe and time as pacing motion as it flows equably through it. SR and GR are even more hostile to Leibniz’s idea of doing away with space and time in favor of relations among particles. SR eradicated the sharp separation between space and time (and with it the notions of absolute simultaneity, distance, duration, velocity, etc.). GR went further by allowing for non-Euclidian space-time and turning space-time from an arena into an active player alongside of matter and fields. It is interesting to note that throughout these changes the idea of an inertial path inherited from Newton (who inherited the idea of natural motion from the Greeks) remains central in the formulations of laws of motion.

5 Quantum Mechanics

The development of the atomic theory of matter that at first seemed to support the Newtonian idea that the ontology of the world consists of minute particles then proved to be its undoing. The central problem again concerned the interaction between charged matter and radiation. It was discovered that atoms consist of positively charged particles surrounded by much lighter negatively charged particles. Electrodynamics and mechanics imply that this configuration is not stable. Negatively charged particles would be attracted by the positively charged particles and fall into the atom’s nucleus emitting radiation. In other words, matter would collapse and the world would come to an end! But this was just the tip of the iceberg. During the first decades of the twentieth century an enormous amount of evidence accumulated involving the behavior of charged particles, atoms, and light that was flatly incompatible with Newtonian mechanics and electrodynamics. It seemed that in some circumstances atomic particles behaved like waves but in other circumstances like particles depending, it seemed, on what observations an experimenter chose to make. This behavior is extremely puzzling calling into question the possibility of a realist understanding of the world.

In response to these problems physicists developed a theory, Quantum Mechanics (QM) that avoids the disastrous consequences of Newtonian

mechanics and electromagnetism and correctly predicts the vast amount of data concerning atoms and radiation. Here are the basics of how elementary QM works. A system S is fully characterized by a mathematical object $\Psi_s(t)$ —the wave function of S at t —which is a vector in an appropriate Hilbert space. $\Psi_s(t)$ replaces the Newtonian notion of the state of a system (the positions and momenta of all the particles that constitute the system). $\Psi_s(t)$ specifies that some (but only some) of the properties traditionally associated with particles have determinate values and what those values are and also specifies the probabilities of the outcomes of measurements of any property relevant to the S . When $\Psi_s(t)$ assigns a determinate value to a property O then $\Psi_s(t)$ is said to be “an eigenstate” of O . Curiously when $\Psi_s(t)$ is an eigenstate of some property O (e.g. position) it is not an eigenstate of certain other properties O^* (e.g. momentum) although it does predict the probabilities of outcomes of measurements of O^* . This is the gist of Heisenberg’s uncertainty principle.

More generally, if $\Psi_{sk}(t)$ is an eigenstate of the position of an electron located at point k and $\Psi_{sk'}(t)$ is an eigenstate of the electron being located at k' then there is a state $a\Psi_{sk}(t) + b\Psi_{sk'}(t)$ called the *superposition* of these states. An electron in this state lacks (if this is the complete physical state) a determinate position. However the coefficients a and b specify the probabilities of finding the electron at k and at k' when position is measured. One of the most notable features of QM is the existence of so-called entangled states that involve particles in distinct regions of space. There are entangled states of a pair of particles states that are eigenstates of the outcomes of measurements of certain properties of the pair being correlated but are not eigenstates of any of these properties on either electron. In other words, QM (on the usual understanding) is saying that the values of the properties are correlated but there is no matter of fact (until measurement) as to what those values are.

QM replaces Newtonian dynamics with a linear deterministic equation specifying the time evolution of $\Psi_s(t)$ (Schrödinger’s equation). This law holds in all situations *except* when measurements are made. If a measurement of a property quantity O (say position) of a system is measured then $\Psi_s(t)$ “collapses” into a new wave function $\Psi_s^*(t^*)$ in which O has a determinate value. The probability associated with the collapse is the same as the probability that a measurement of O on S at t yields that determinate value. In other words, if an electron is in state $a\Psi_{sk}(t) + b\Psi_{sk'}(t)$ and its position is measured the collapse will result in a state in which the electron is located at k, k' with probabilities associated with a and b .

The wave function of an atom of hydrogen (consisting of an electron “orbiting” a proton) ensures the stability of the atom, appropriate wave functions predict the frequencies of light emitted by excited helium atoms, predict the periodic table, and so on. As a predictive tool QM is enormously successful. But even my brief description of the theory should be enough to see that it is

very puzzling. What reality can possibly lie behind the wave function? What can be going on when a measurement is made? One reply associated with the orthodox or “Copenhagen” interpretation is that one should not press these questions very hard and instead should be content with an instrumentalist understanding of the QM.

The appeal of the Copenhagen view is that if one tries to understand QM in a realist way one is confronted with problems at every turn. First off the idea of a particle possessing a determinate position but not determinate velocity is, well, mind boggling (as is there being a determinate correlation but no determinate values that are correlated). Note, QM is not saying that when $\Psi_s(t)$ assigns a determinate position to S then S ’s velocity is 0. Rather, if $\Psi_s(t)$ is the complete state of S (as the standard account of QM maintains) then it says that S has no determinate velocity at all! How can it be that an electron is in a state in which it has no determinate position but assigns a probability to finding it at a particular location? Further, if the only dynamics of the wave function were Schrödinger’s linear equation and if QM applies to macro systems like cats then, as Schrödinger showed, there are easily describable situations in which the wave function applicable to the cat fails to assign to the cat a state in which it is determinately alive or determinately dead.

The *measurement problem* in QM is the problem that in measurement (and many other) interactions if the governing law is expressed by Schrödinger’s equation then the post measurement state of the measuring apparatus + system measured does not specify an outcome. The “collapse dynamics” was introduced exactly to avoid this absurd (and self-defeating) consequence. One cost of “solving” the measurement problem this way is that there are two radically different laws of evolution for quantum states; one deterministic and linear and the other probabilistic and nonlinear. The cost is steep since this way amounts to characterizing the fundamental laws in terms of a vague macroscopic notion of “measurement.” Exactly which interactions count as measurements? One idea that was seriously proposed is that measurements occur only when a conscious being interacts with the physical system. It was suggested that mind is required for the existence of a determinate physical world. It is not surprising then that many physicists are content to construe QM as a mere predictive instrument and that some have even claimed that the success of QM demonstrates that a realist theory of a mind-independent world—the kind of theory that Newton, Maxwell, Boltzmann and Einstein were hoping to find—is out of the question.

Or is it? Einstein noticed that the anti-realism of the orthodox interpretation was closely connected to its reliance on the collapse and the assumption that the quantum mechanical description of state $\Psi_{sk}(t)$ is complete; that is, there are no further facts about the situation of, for example, an electron whose QM state is $\mathbf{a}\Psi_{sk}(t) + \mathbf{b}\Psi'_s(t)$ that would determine that it is located at k or that it is

located at k' . Proponents of the Copenhagen interpretation gave proofs—no hidden variable proofs—that they thought showed that the QM state cannot be supplemented by further facts.³⁶ To the contrary, Einstein famously gave an argument that he thought showed that the QM state cannot be the complete state of the electron. Briefly, his argument (in a recent version due to Bohm) is this:

There is a family of infinitely many properties of electrons (“spin properties”) each of which can take one of two values—“up,” “down”—such that when the state of a single electron is an eigenstate of one of these properties it is not an eigenstate of any other property in the family. Also there is an entangled state (the EPR [Einstein-Podolsky-Rosen] state) of a pair of electrons which is not an eigenstate of any of the spin properties for either electron. However, this state is an eigenstate of the values of the properties (for any of the properties) being correlated. In other words if a measurement of one of these properties is made on one of the electrons and the result is “up” then a measurement of the same property on the other one will yield “down.” Einstein imagined a situation in which the electrons are spatially separated (it makes no difference how far apart) and one of the properties P of one of the electrons is measured. According to orthodox QM upon measurement of P the state of the pair of electrons instantaneously collapses into a state in which both electrons have determinate values of P . Einstein noticed that this collapse seemed to involve some kind of non-local influence since the state of both electrons was altered by a local interaction with just one. He reasoned that the non-locality could be avoided only if the two electrons had determinate values of P all along that were revealed by and not brought about by measurement.³⁷ But this would mean that the QM state is not the complete state. In other words, he seemed to suggest that if and only if there is more to the state of a system than its wave function described then the collapse law with its “spooky” non-locality and the concomitant instrumentalism could be avoided.

Bohr and Pauli (and other defenders of the orthodox account) were not impressed by Einstein’s argument or at least did not budge from their claim that the QM state is complete. But 15 years later David Bohm devised an alternative to QM in which particles always have positions and the quantum state is not merely a mathematical object but represents something like a force field that guides the motion of the particles associated with it. The time evolution of this field is governed only by the linear Schrödinger law (no collapse) and there is an additional law (the guidance equation) that determines how the field guides its respective particles. In order to recover the probabilistic

predictions of QM Bohm also added a probability distribution over possible locations of particles that is also determined by their quantum state in accordance with the usual QM prescription. Bohmian probabilities play a role not unlike the statistical mechanical probabilities.³⁸

Bohmian mechanics makes the same probabilistic predictions as orthodox QM for the outcomes of measurements that are recorded in the positions (as all measurements ultimately are).³⁹ But unlike orthodox QM there is no “collapse” of the quantum state and the notion of measurement plays no role in the formulation of the theory. The apparent “collapse” of the wave function in measurement is explained by the theory itself. Schrödinger’s cat paradox (and more generally the measurement problem) is solved since although the quantum state does not specify whether the cat is alive or dead the positions of its particles do.

Bohmian mechanics can be understood as an account of the fundamental ontology and laws of a mind-independent world. One might think, at first, that this is just the sort of theory Einstein was hankering after since it rejects the completeness of the quantum state and adds to the state positions of particles. Also its dynamical laws are deterministic (no dice playing) to boot. However, Einstein did not embrace it. Perhaps the reason was the way it handles the EPR experiment. On the Bohmian account the outcomes of the measurements of spin properties are determined by the EPR quantum state, and the exact positions of the electrons prior to the measurements. However, it turns out that the way the quantum state yields correlated results is for the *first* measurement made on one of the electrons to alter the quantum state in just such a way as to “guide” the second electron into a trajectory that guarantees the correlation. In other words Bohmian mechanics has nonlocality built right into it. Einstein would have seen trouble for relativity on this account.

The way the Bohmian laws work there needs to be a matter of fact about which measurement of the electrons occurs *first* (if the other electron had been measured first the outcomes of both measurements might have been different.) This means that there is a conflict between Bohmian mechanics and SR understood as claiming that Minkowski space-time is the whole structure of space-time since there may be no matter of fact about which measurement occurs first. Bohmian mechanics requires there to be a preferred reference frame that is a real part of the space-time just as Lorentz assumed. So Bohmian mechanics was definitely not what Einstein had in mind when he imagined a replacement for QM.⁴⁰

The nonlocality of Bohmian mechanics and its reliance on a preferred reference frame are often brought up as objections to it by adherents to the orthodox approach. But it should be recalled Orthodox QM (the collapse law) is also nonlocal and also has trouble with relativity.⁴¹ An instrumentalist perspective

encourages proponents of the orthodox account to not be much troubled by the mismatch between QM and relativity. A more serious problem is that the Bohmian account has not yet been successfully developed for quantum field theory. In any case, Bohm's account did not receive a positive reception from the physics community and only fairly recently has it begun to be seen by some physicists and philosophers of physics as a promising realist alternative to orthodox QM.⁴²

The question naturally arises as to whether the nonlocality found in Bohmian mechanics and in orthodox QM is inevitable. Could there be a satisfactory theory of the world that yields the same predictions as QM but is local?⁴³ In 1964 John Bell (Bell 2008) produced a simple proof that demonstrates that *no* local theory can recover exactly the predictions of QM. Further, the relevant predictions of QM that implicate nonlocality have been tested and found to be correct. The astonishing conclusion is that nonlocality is a feature of our world.⁴⁴ In other words, an event occurring in region *R* of space-time can affect what happens in some other region *R** even when these regions are so separated that a light signal (or any other physical process) can get from *R* to *R**. This may be the most important consequence of recent physics for metaphysics. One the face of it nonlocality is a problem for various traditional metaphysical doctrines.⁴⁵

There are other "interpretations" of QM in addition to Bohmian mechanics that can be understood as attempts to provide realist accounts. The most significant is Everett's "Many Worlds" interpretation. The basic idea of the Many Worlds theory is that the measurement by a macroscopic measuring device of a particle that is in a quantum state like $a\Psi_{sk}(t) + b\Psi'_s(t)$ (a superposition of the electron being located at *k* and being located at *k'*) results in the "branching" of the universe into two distinct, noninteracting universes, in one of which the particle is measured to be at *k* and in the other it is measured to be at *k'*. There are a number of ways of developing the Everett account. One is to take the basic ontology of the universe to be an evolving (in accordance with the Schrödinger law) quantum state understood as a kind of field that inhabits a very high (perhaps infinite) dimensional space. As it evolves the wave function "decoheres" with respect to certain degrees of freedom of the field and these "bunch up" and evolve as though they were separate "worlds."⁴⁶ The Everett account faces two big issues. One, which has received a great deal of attention recently, is making sense of probabilities in the account. The Schrödinger evolution is deterministic and unlike Bohm's theory there are no extra matters of fact (fundamental particle positions) over which probabilities can be defined. The other problem is explaining how our material world (perhaps along with many other worlds) with its quotidian material objects supervenes on a branch of the quantum state.⁴⁷

6 Conclusion

At the beginning of the twentieth century physicists were faced with the problem of reconciling Newtonian mechanics and electromagnetic theory. The results were relativistic space-times and quantum mechanics. At the start of the twenty-first century physicists are faced with the problem of reconciling these. We already saw that realist versions of quantum mechanics seem to require a preferred reference frame and so to that extent are already at odds with SR. But the conflict between QM and relativity goes much deeper. Reconciliation of the two will require a quantum mechanical account of gravitation. While there are ideas about how that might be accomplished (string theory, loop quantum gravity) it is an open problem. Instrumentalists may be content to apply GR to the very large and QM to the very small and not worry so much about reconciling the two. But the realist dream (the dream of Einstein) will be realized only by a complete theory that specifies space-time, the particles and fields, etc. that inhabit it and the laws that govern the two. However that will go, it is certain that there is a lot in physics and philosophy of physics to be digested by metaphysics and epistemology.

Notes

- 1 The idea that the primary source for metaphysics should be physics has been much discussed and forcefully advocated in recent years (Maudlin 2007, Ladyman and Ross 2007, Paul 2012). Of course the idea is not new (Descartes, Leibniz, and Newton among many others adhered to this view) but a good deal of metaphysical speculation and debate in the twentieth century has proceeded without much attention to developments in physics, especially developments involving relativity theories and quantum mechanics. The importance of the philosophy of physics to other parts of physics and the need for philosophers to have an acquaintance with physics and the philosophy of physics is emphasized by Michael Dummett; “The greatest lack, however, is of philosophers equipped to handle questions arising from modern physics [or older physics!]; very few know anything like enough physics to be able to do so. This is a serious defect, because modern physical theories impinge profoundly upon deep metaphysical questions it is the business of philosophy to answer. We may hope that some philosophers may become sufficiently aware of this lack to acquire a knowledge of physics adequate both to integrate it with their treatment of metaphysical problems and to convey to philosophical colleagues who know less physics what they are talking about” (Dummett 2010).
- 2 In standard discussions of mechanics a particle’s velocity is defined as the time derivative of the particle’s position. But there are two views concerning the metaphysical nature of velocity. One is that a particle’s velocity at each time is metaphysically derivative on its trajectory. The other is that its trajectory is metaphysically derivative on its velocity at each time. Given Newton’s notion of time as “flowing” the second seems more apt.
- 3 $F(1,2)$ is the gravitational force that particle 1 exerts on particle 2 and $r(1,2)$ is the distance between the two particles. The direction of the force particle 1 exerts on particle 2 is toward particle 1 and visa versa. $F(2,1) = -F(1,2)$.

- 4 This is an oversimplification. There are solutions to Newtonian equations (depending on what velocities and forces are allowed) that violate determinism but involve special initial conditions and forces. See Earman 1986.
- 5 Newtonian mechanics is often presented as though forces are some kind of entity or power and so are among the world's primitive ontology. It is possible to formulate mechanics without any reference to forces and simply in terms of equations of motion of particles, for example, Hamilton's equations. It is not clear whether Newton thought of laws as regularities grounded in forces and material particles or as something like divine edicts that govern particles and forces or in some other way.
- 6 Of course points in absolute space already violate the PII.
- 7 Scholium to Definition 8 of the *Principia* in Newton 1934.
- 8 A "Galilean" transformation takes coordinates of one inertial frame to the coordinates of another inertial frame by the mapping $x'=x-vt$, $y'=y$, $z'=z$, $t'=t$.
- 9 Newton describes a thought experiment similar to this one (Newton 1934, vol. 1, p. 12.). This version is due to David Albert.
- 10 But relationist mechanics is almost deterministic since the state of the spring and balls at $t(0)$ and a time close to $t(0)$ will determine the extent of rotation.
- 11 For recent attempts to carry out Mach's program see Barbour 1999.
- 12 For example, the relationist theory can be the claim that a history of changes in the distances between particles is physically possible if, and only if, these distances can be embedded in a fictional absolute space in such a way as to satisfy the Newtonian Laws. So absolute space is employed as part of a fiction to describe the actual world. This approach is not likely to please a realist about laws.
- 13 Barbour 1999.
- 14 The metaphor of a flow of time suggests that there are irreducible tensed facts and so-called growing block or moving spotlight accounts of time (Callender 2012). Whatever these come to they do not seem required by Newtonian mechanics or the other fundamental theories we will discuss.
- 15 The entropy of a macro condition M is given by $S_B(M(X)) = k \log |\Gamma_M|$ where $|\Gamma_M|$ is the volume (on the measure) in Γ associated with the macro state M , and k is Boltzmann's constant. S_B provides a relative measure of the amount of Γ corresponding to each M . Given a partition into macro states the entropy of a micro state relative to this partition is the entropy of the macro state that it realizes.
- 16 So the second law should not have been stated in the first place as an absolute prohibition on the entropy of a system decreasing but rather as being enormously unlikely.
- 17 If the Boltzmann probability posit is applied to the macro condition of the universe at t since it implies that it is likely that this macro condition arose out of higher entropy states and in particular this means that the "records" in books, etc. likely arose out of chaos and not as accurate recordings of previous events. This undermines the claim that there is evidence reported in those books that supports the truth of the dynamical laws and so results in an unstable epistemological situation.
- 18 Also, the prescription will prescribe incompatible probabilities at different times since the uniform distribution over the macro state at t will differ from the uniform distribution over the macro state at other times.
- 19 See Sklar 1993 for a discussion of some proposals for responding to the reversibility paradox.
- 20 Although there are issues concerning how to think of entropy in the very early universe it is generally held that cosmology supports the claim that right after the big bang the entropy of the universe was very tiny. This may strike one as counterintuitive since at the big bang the universe was enormously tiny and dense with matter/

energy uniformly distributed in space. But because gravitation acts to clump matter this is a very low entropy condition. For a discussion see Callender 2010, Carroll 2010, Penrose 2005, and Greene 2004.

- 21 This idea is not original with Albert. For example, it is explicit in a lecture by Feynman 1994.
- 22 While the account is developed on the assumption of Newtonian mechanics the same considerations carry over to deterministic versions of quantum mechanics (e.g. Bohmian mechanics, and Everettian QM). If the dynamical laws are probabilistic (as on GRW theory) then the initial probability distribution may no longer be needed although the past hypothesis still plays the role it plays in the account that I sketch. See Albert 2000 for a discussion.
- 23 Maudlin suggested in discussion that if the uniform probability distribution accomplishes all Albert claims for it then infinitely many other distributions will do as well. If so and if probabilities are understood objectively in the way I discuss later then there may be empirically discernable differences among these distributions or it may be a case of massive underdetermination. It is reasonable to posit the uniform distribution since it is the simplest until evidence is adduced against it.
- 24 It is thought that the length of time it would take for entropy to increase to equilibrium is far greater than the approximately 14 billion years that have passed since the Big Bang.
- 25 See Albert 2000 for a discussion of how a Maxwell demon may prepare a system so that its entropy likely decreases.
- 26 See Albert 2000 and Loewer 2012 for suggestions of how the statistical mechanical account may explain all of time's arrows.
- 27 There are discussions of probability in statistical mechanics and in quantum mechanics in Loewer 2001 and 2004.
- 28 Roger Penrose (1989) has been especially insistent on this point. He estimates that the probability of this particular type of past state occurring is 1 out of $10^{10^{23}}$.
- 29 For a discussion of these points see Loewer 2012, Carroll 2010, and Callender 2003. The theological proposal is an example of finetuning argument for design and a designer.
- 30 Of course this is immensely anachronistic. No philosopher I know of looked at physics at the time in quite this way and if she did she would have heard a lot of conflicting answers among which are those in my list.
- 31 The Lorentz transformations are: $x' = (x - ut) / \sqrt{1 - u^2 / c^2}$, $y' = y$, $z' = z$, $t' = \{t - (u / c^2) x\} / \sqrt{1 - u^2 / c^2}$.
- 32 See Bell 2008 for an account of Lorentz's proposal.
- 33 Metaphysicians who think that shape at t is an example *par excellence* of an intrinsic property should take note.
- 34 This consequence of relativity explains why cosmic rays entering the earth's atmosphere at high speed live longer before decaying than would the same particles at rest.
- 35 See Carroll 2010 for an accessible discussion of the physics of time travel.
- 36 There is a famous argument due to von Neumann that allegedly establishes this. For a discussion of this argument and how it goes wrong see Bell 2008.
- 37 This account is rough but gives the gist of the argument. For a careful exposition see Bell 2008.
- 38 But the wave function in Bohm plays a very different role from the macro state in statistical mechanics since the former is an element of fundamental ontology.
- 39 This claim requires some qualification. Since orthodox QM specifies that the quantum state collapses in measurement interactions there will be some in principle differences in the predictions it makes and the predictions Bohmian mechanics make

- regarding the measurements of certain very complex properties of the measurement apparatus. But we will never be in a position to know exactly what properties these are or even if we did make such measurements. So for all practical purposes orthodox QM and Bohmian mechanics are empirically equivalent.
- 40 Einstein said of Bohm's theory that "it is too cheap." Exactly what he had in mind is not completely clear but he was likely bothered by its apparent conflict with SR (Cushing 1994).
 - 41 The collapse can be formulated in Minkowski space-time (Aharonov and Albert 1984) but it is difficult to take their proposal as a serious realist account of the evolution of the quantum state.
 - 42 Bell (2008) argued that Bohm's account should be taken seriously and subsequently it has been greatly developed and clarified by Shelly Goldstein and his group (Goldstein 2012). In philosophy of physics one hardly sees a mention of Bohm's theory prior to the books by Albert (1992) and Maudlin (1994).
 - 43 Defining what it is for a theory or law to be local is subtle. The basic idea is that if A affects B it does so via a chain of events that are in an appropriate sense next to each other in 3-D space and time. Bohmian mechanics clearly violates this for the causes of an event in a small spatial region R. For a discussion on this see Maudlin 1994.
 - 44 There have been a number of attempts to get around Bell's argument and the empirical evidence that implicates nonlocality. Suffice it to say that in my view non are successful.
 - 45 Entangled states and the resulting nonlocality of QM are incompatible with the view that a complete account of the universe is specifiable by the contents of each small region of space-time (David Lewis' doctrine of Humean Supervenience). See Lewis 1987, Loewer 1996, and Maudlin 2007.
 - 46 Decoherence of the wave function depends on the particular wave function of the system and its environment and the law of evolution.
 - 47 For discussions of Many Worlds theories see Albert and Loewer 1988 and Saunders et al. 2010, and Wallace 2012.

Bibliography

- Aharonov, Y. and Albert, D., 1984. "Is the Usual Notion of Time Evolution Adequate for Quantum Mechanical Systems? II Relativistic Considerations." *Physical Review*, D 29, pp. 228–34.
- Albert, D., 2000. *Time and Chance*. Cambridge, MA: Harvard University Press.
- , 1992. *Quantum Mechanics and Experience*. Cambridge, MA: Harvard University Press.
- Albert, D. and Loewer, B., 1988. "Interpreting the Many Worlds Interpretation." *Synthese*, 77 (2), pp. 195–213.
- Barbour, J., 1999. *The End of Time*. Oxford: OUP.
- Bell, J. S., 2008. *Speakable and Unspeakable in Quantum Mechanics*. 2nd edn. Cambridge, UK: CUP.
- Bohm, D., 1952b. "A Suggested Interpretation of the Quantum Theory in Terms of 'Hidden Variables' II." *Physical Review*, 85, 180–93.
- , 1952a. "A Suggested Interpretation of the Quantum Theory in Terms of 'Hidden Variables' I." *Physical Review*, 85, 166–79.
- Callender, C., 2012. "Time's Ontic Voltage." In Adrian Bardon, ed., *The Future of the Philosophy of Time*. New York: Routledge, pp. 73–98.

- , 2010. "The Past Hypothesis Meets Gravity." In Gerhard Ernst and Andreas Hüttemann, eds, *Time, Chance and Reduction. Philosophical Aspects of Statistical Mechanics*. Cambridge: CUP, pp. 34–58.
- , 2003. "Is there a Puzzle about the Low Entropy Past?" In C. Hitchcock, ed., *Contemporary Debates in the Philosophy of Science*. Oxford: Blackwell, ch. 12.
- Carroll, S., 2010. *From Eternity to Here: Quest for the Ultimate Theory of Time*. New York: Dutton.
- , 2004. *Spacetime and Geometry: An Introduction to General Relativity*. Boston, MA: Addison-Wesley.
- Cushing, J. T., 1994. *Quantum Mechanics: Historical Contingency and the Copenhagen Hegemony*. Chicago: University of Chicago Press.
- Dummett, M., 2010. *The Nature and Future of Philosophy*. New York: Columbia University Press.
- Earman, J., 1995. *Bangs, Crunches, Whimpers, and Shrieks: Singularities and Acausalities in Relativistic Space-Times*. Oxford, UK: OUP.
- , 1989. *World Enough and Space-Time*. Cambridge, MA: MIT Press.
- , 1986. *A Primer on Determinism*. Dordrecht, NL: Reidel.
- Feynman, R. P., 1967. *The Character of Physical Law*. Boston, MA: MIT Press.
- Feynman, R. P., Leighton, R., and Sands, M., 1975. *The Feynman Lectures on Physics*. Reading, MA: Addison-Wesley.
- Friedman, M., 1986. *Foundations of Space-Time Theories: Relativistic Physics and Philosophy of Science*. Princeton, NJ: Princeton University Press.
- Geroch, R., 1978. *General Relativity from A to B*. Chicago, IL: University of Chicago Press.
- Goldstein, S., 2012. "Bohmian Mechanics." *Stanford Encyclopedia of Philosophy* (Fall edn). E. N. Zalta, ed. Available at: <http://plato.stanford.edu/archives/fall2012/entries/qm-bohm/>
- Greene, B., 2004. *The Fabric of the Cosmos: Space, Time, and the Texture of Reality*. New York: Knopf.
- Horwich, P., 1987. *Asymmetries in Time*. Cambridge, MA: MIT Press.
- Huggett, N., ed., 1999. *Space from Zeno to Einstein*. Cambridge, MA: Bradford.
- Ladyman, J. and Ross, D., 2007 *Everything Must Go: Metaphysics Naturalized*. Oxford: OUP.
- Lewis, D. K., 1987. *Philosophical Papers*. Oxford: OUP.
- Loewer, B., 2012. "Two Accounts of Laws and Time." *Philosophical Studies*, 160 (1), pp. 115–37.
- , 2002. "David Lewis' Humean Theory of Objective Chance." *Philosophy of Science*, 71 (5), Proceedings of the 2002 Biennial Meeting of The Philosophy of Science.
- , 2001. "Determinism and Chance." *Studies in the History and Philosophy of Science*, 32 (4), pp. 609–20.
- , 1996. "Humean Supervenience." *Philosophical Topics*, 24, pp. 101–27.
- Maudlin, T., 2012. *Space and Time*. Oxford: OUP.
- , 2007. *The Metaphysics Within Physics*. Oxford, UK: OUP.
- , 1994. *Quantum Non-Locality and Relativity: Metaphysical Intimations of Modern Physics*. Cambridge, MA: Blackwell.
- Misner, C., Thorne, K., and Wheeler, J., 1983. *Gravitation*. San Francisco, CA: W. H. Freeman.

- Newton, I., 1999. *The Principia: Mathematical Principles of Natural Philosophy*, (trans.) I. B. Cohen and A. Whitman. Berkeley, University of California Press.
- , 1934. *Principia*. Translated by Andrew Motte, Revised by Florian Cajori. 2 vols. Berkeley, CA: University of California Press.
- Penrose, R., 2005. *The Road to Reality*. New York: A. E. Knopf, ch. 27–9.
- Saunders, S., Barrett, J., Kent, A., and Wallace, D., eds, 2010. *Many Worlds? Everett, Quantum Theory, and Reality*. Oxford: OUP.
- Sklar, L., 1993. *Physics and Chance. Philosophical Issues in the Foundations of Statistical Mechanics*. Cambridge: CUP.
- , 1977. *Space, Time, and Spacetime*. Berkeley, CA: University of California Press.
- Wallace, D., 2012, *The Emergent Multiverse: Quantum Theory According to the Everett Interpretation*. Oxford: OUP.

21 Causation

Helen Beebee

1 Introduction

The concept of causation pervades our ordinary talk and thought about the world. We routinely want to know what caused a given phenomenon, whether it is a car crash, a broken washing machine or a friend's bad mood, or what the likely effects of a given phenomenon will be. (What effect on house prices would a hike in interest rates have? Would a trip to the cinema improve my friend's mood?) And—as Elizabeth Anscombe (1971) famously pointed out—very many transitive verbs enshrine the concept of causation: to break something is to cause it to be in a broken state, to hurt someone is to cause them pain, and so on.

The concept of causation has also—at least in the last half-century or so—pervaded philosophical theorizing. There are, for example, causal theories of knowledge (roughly: to know that p is to have been caused to believe that p by the obtaining of the fact that p ; see e.g. Goldman 1967), perception (roughly: to perceive that o is to have been caused to have an experience of o by the presence of o ; see e.g. Grice 1961), rational decision-making (roughly: rational decision requires one to choose one's action on the basis of what the desirability of the likely effects of that action; see e.g. Lewis 1981), moral value (the goodness or badness of an act is determined by its consequences), and so on.

The concept of causation therefore seems to be an extremely important one (though, as we shall see, this claim has been disputed). But what is it that we attribute to a pair of events—my dropping a cup, say, and its breaking—when we claim that the first is a cause of the second? For it to be true that a caused b , it needs to be true not merely that a and b both happened, or that a happened (just) before b . Just after I dropped the cup, the telephone rang—but *that* was not caused by my dropping the cup. So what more is needed? As a first pass, we might try to say that a and b need to be *connected* in some way: perhaps we might say that a *produced* b , or *made* b *happen*, or *brought* b *about*. But it seems that we are really just trading synonyms here; none of these ways of characterizing the causal connection sheds any real light on its nature in the absence of further analysis.

This central question concerning its nature has dominated the extensive literature on the topic of causation in analytic philosophy in the last 50 years or so, and Sections 3 and 4 of this chapter explore some of the answers that have been proposed. In Section 2, I set the context for these answers with a very brief and partial account of two earlier contributions to the debate by David Hume and Bertrand Russell. In Section 3, I briefly rehearse one of the central points of dispute, namely whether causation should be seen in a “Humean” or “anti-Humean” light. In Section 4, I survey some of the most frequently discussed kinds of theory of causation. Finally, in Section 5 I discuss a selection of related issues concerning causation. What exactly does the relation of causation relate? (Events? Facts? Objects? All of the above?) Can absences or omissions be causes? Is the concept of causation univocal, or is there really more than one concept of causation? I also briefly discuss the “causal exclusion problem.”

2 Hume and Russell

As an empiricist, Hume holds that every “idea” (roughly: concept) has its source in an “impression”—an experience of some kind. Having understood the importance of the idea of causation—it is, he says, the only relation that allows us to “go beyond what is immediately present to the senses, either to discover the real existence or the relations of objects” (1739–40, 73)—he spends a large part of Book I of his *Treatise of Human Nature* searching for its elusive impression-source. He identifies three core components of the idea of causation, namely priority (causes precede their effects), contiguity (causes and effects are contiguous, or right next to each other, in space and time) and “constant conjunction” (causes and effects are constantly conjoined, that is if *a* causes *b*, then whenever an event similar to *a* occurs, it is followed by an event similar to *b*). But, he insists, these three conditions are not enough: our idea of causation includes the idea of “necessary connection.” After all, if the relation of causation is to inform us “of existences and objects, which we do not see and feel”—that is to say, if it is to underpin our *inferences* from causes to effects (as when I drop the cup and infer that it will shortly break)—then it would seem that causes must *guarantee* the occurrence of their effects, or, in other words, they must *necessitate* their effects. And so the impression-source of the idea of necessary connection must be found if the idea of causation is to be legitimized.

As it turns out, Hume eventually turns this last thought on its head: rather than our inference from cause to effect being underpinned by our responding to an objective necessary connection between the two, it turns out that the inference itself is the impression-source of the idea of necessary connection.

Once I have experience of *As* being constantly conjoined with *Bs*, on observing an *A* I infer that a *B* will follow—not through some sophisticated piece of reasoning, but through “Custom or Habit” (1748/1751, 43), or by what psychologists nowadays call “associative learning.” So there is a “transition” in the mind from the experience of an *A* to the expectation that a *B* will follow. This transition itself produces an impression, and that impression, Hume claims, is the source of the idea of necessary connection. Contrary to what we might have thought, then, the necessity that is part and parcel of our concept of causation is supplied not by the world, but by our own minds; indeed, the supposed notion of a “nexus” between cause and effect that is independent of the mind is unintelligible, since it lacks an impression-source.

For present purposes, Hume’s account of causation raises two central and closely connected sets of questions. First, does causation somehow or other involve *necessitation*—do causes *make* their effects happen? If so, what is the nature of this necessity? Is it something supplied by the mind, or is it instead a feature of mind-independent reality—and, if the latter, is such a thing genuinely intelligible? Second, does causation require “constant conjunction” or *regularity*? Is it essential to our concept of causation that the same cause will always and everywhere be followed by the same effect? If so, can this fact be used to shed light on the nature of causation?

These two sets of questions are taken up by Bertrand Russell in his “On the notion of cause” (1912–13). Russell begins by complaining that “the word ‘cause’ is so inextricably bound up with misleading associations as to make its complete extrusion from the philosophical vocabulary desirable” (1912–13, 1), and, famously, noting that the “law of causality”—roughly, the principle of “same cause, same effect”—“is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm” (ibid.). Russell broadly agrees with Hume on the unintelligibility of the notion of some mind-independent nexus between causes and effects. But he is equally dismissive of Hume’s positive contention that causation requires contiguity, temporal priority, and constant conjunction. In particular, he argues that the regularities we observe in daily life are almost all merely “fairly dependable” rather than exceptionless (1912–13, 8); and once we appeal to the sciences to supply our supposed exceptionless regularities, what we find is that once we have specified the full cause of a given event in sufficient detail to render our regularity exceptionless, we’ll find ourselves with a circumstance that is unlikely ever to be repeated and hence cannot be considered to be an instance of a “regularity” at all.

Perhaps the most enduring aspect of Russell’s critique of the notion of cause, however, is his claim that “in advanced sciences such as gravitational astronomy, the word ‘cause’ never occurs” (1912–13, 1). For Hume, the relation of causation is central to our ability to draw inferences beyond what we

can “see or feel”; for Russell, our best science, namely physics, has “ceased to look for causes” because “there are no such things” (ibid.); and physics is none the worse—indeed it is all the better—for that.

3 Humeanism and Its Critics

Let’s return to the two sets of questions posed above, concerning the idea that causes necessitate their effects and the idea that causation and regularity are intimately related. One way to answer these questions—and this is a view that has been attributed to Hume by many commentators—is to claim that causation just *is* a matter of regularity: for *a* to cause *b* just *is* for all events like *a* (the *As*) to be followed by events like *b* (the *Bs*). So *a* causes *b* if and only if all *As* are followed by *Bs*. This view is sometimes called the “naïve regularity theory” of causation. According to the naïve regularity theory, there is no real necessity involved in causation at all. Or, to put it another way, while it is true that “all *As* are followed by *Bs*” and “an *A* occurs” entail that a *B* will follow, and entailment is a species of necessity, the “necessary connection” between *a* and *b* obtains merely in virtue of the fact that all *As* are followed by *Bs*: there is no intrinsic relation of necessity, or indeed a nexus of any kind, between *a* and *b*.

Unfortunately the naïve regularity theory is subject to an insurmountable battery of objections, of which I shall mention just two: the problem of accidental regularities and the problem of the common cause. There are some truths of the form “all *As* are followed *Bs*” where it is manifestly not the case that *As* cause *Bs*. Starting with the problem of accidental regularities: imagine that on all four occasions when I took a bath in my previous house (I very rarely take a bath) (*A*), the neighbor rang the door bell (*B*) just as I stepped in. So all *As* have been followed by *Bs* so far, and since I have now moved house, all the *As* there will ever be are followed by *Bs*. Of course, it’s *possible* that the neighbor somehow knew I was taking a bath and deliberately rang the bell to annoy me—in which case, my taking a bath *would* have been a cause of her ringing the bell. But according to the naïve regularity theory, my taking a bath was automatically a cause of the neighbor’s ringing the bell, just in virtue of the fact that the former was always followed by the latter. Clearly that’s not right: it’s just an accident that the regularity holds. The problem of the common cause is similar. Imagine that every time a particular barometer points to “rain,” it rains shortly afterward. According to the naïve regularity theory, the position of the barometer pointer causes the rain. But manifestly this is not so: the position of the barometer pointer and the rain are effects of a common cause, namely low air pressure.

Despite the obvious failings of the naïve regularity theory, many philosophers have attempted to provide more sophisticated “regularity theories” of

causation, which seek to ground facts about causation in facts about regularities but conceive the relationship in a more complex way that avoids the problems that beset the naïve version (see e.g. Mackie 1965). The version of the regularity theory that has received the most attention, and perhaps the one with the best chance of success, is the counterfactual analysis of causation first developed by David Lewis (1973a) and discussed in Section 4.

Regularity theories of causation are often referred to as “Humean” theories not just because they preserve the spirit of the naïve regularity theory that is often attributed to Hume, but also for the deeper reason that they eschew the kind of nexus or mind-independent necessity that Hume claimed to be unintelligible. The eschewal of necessity that has driven some philosophers in the direction of a regularity theory also no doubt gained some momentum from W. V. Quine, according to whom “*de re*” necessity (that is, all necessity that does not ultimately derive from the meanings of words, as in “necessarily, all bachelors are unmarried”) is deeply suspicious and therefore—echoing Russell—has no place in philosophical discourse (Quine 1963).

It is worth remembering, however, that Hume himself—unlike contemporary regularity theorists—apparently endorses the claim that the idea of necessity is an essential part of the idea of causation; it’s just that the necessity involved in causation turns out contributed by the mind rather than the world (see Beebe 2006, chs.4 and 7). Other kinds of broadly Humean theory stay closer to what may have been Hume’s real view in this regard. For example, Peter Menzies and Huw Price (1993) argue that our concept of causation has its roots in our experience not as predictors, as Hume had it, but as agents. Price (2007) later characterizes causation as a “perspectival” phenomenon. The world exhibits a fundamental asymmetry in that entropy tends to increase over time; and it is because of this asymmetry, Price thinks, that we think of the future as “open” and the past as “fixed.” Thus when we deliberate we hold fixed what has already happened and regard what has yet to happen as up to us, to be determined (in part) by what we decide to do. The deliberative perspective designates the actions we might perform (such as my putting on the kettle) as means and events that lie in the future (such as my getting my desired outcome of a cup of tea), but not events that lie in the past (such as my having had a cup of tea earlier in the day) as ends; and it is in this deliberative perspective that he locates our conception of the world as a world of causes (means) and effects (ends). A related view is held by Jon Williamson (2006), who develops an “epistemic” theory of causation, according to which—again, in broadly Humean spirit—causation is conceived in terms of what beliefs it would be rational for us to adopt for the purposes of predicting, controlling and explaining what goes on in our environment: “the causal relation is characterised by the causal beliefs that an omniscient rational agent should adopt” (ibid., 7), where the rationality of those causal beliefs is secured by the

right kinds of evidential relations rather than their latching onto some mind-independent causal nexus (see also Ramsey 1929).

It is, however, open to question whether we should agree with Hume about the alleged unintelligibility of necessary connections, and plenty of philosophers simply reject Hume's claim, thereby constituting the "anti-Humean" wing of the debate. For example, Adrian Heathcote and David Armstrong (1991) propose an account of causation that derives from Armstrong's (1983) theory of laws of nature. On Armstrong's account of laws, a law of nature is a relation of necessity (N) between universals, so that it is a law that all F s are (or are followed by) G s—where F and G are universals—just in case $N(F, G)$. The basic idea is that what makes it a *law* that all F s are followed by G s, as opposed to its merely being *true* that all F s are followed by G s, is precisely that in the former case, but not the latter, F and G are related by N . (So—to use toy examples—there is no necessary connection between taking a bath and the neighbor ringing the doorbell; but there *is* a necessary connection between being an object with a particular mass m subject to force f and accelerating at rate a : any object possessing the universals m and f *must* accelerate at rate a , where $f = ma$.) Armstrong and Heathcote hold that causation is simply the instantiation of a law, thus conceived. So while the law relation N relates universals F and G , say, that law is instantiated in particular instances or "states of affairs," so that a *particular* object's having F and its having G —those two states of affairs—are themselves related by an instance of N .

More generally, Hume's claim about the unintelligibility of mind-independent necessity is grounded in a version of empiricism that is no longer widely held. As we saw in Section 2, according to Hume all "ideas" must be derived from "impressions"; in other words, to put it crudely, you can't have a concept of something that you have no experience of (unless you can somehow construct the concept out of materials you *do* have experience of—so, for example, we can form the idea of a unicorn by combining the idea of a horse with the idea of a horn, both of which we have experience of). But this version of empiricism is widely rejected as too demanding. In particular it seems to have unpalatable consequences for the kinds of unobservable entity that are commonplace in scientific theories: quarks, electrons, forces, fields, and so on. Since we can have no direct experience of such entities, Hume's empiricism seems to entail that we cannot so much as form concepts of them—which makes it rather hard to see how to make any sense of contemporary physics. These days, many philosophers count themselves as empiricists in a more relaxed sense that allows us to have legitimate concepts of unobservable entities, so long as those entities have a clearly definable role within the theoretical framework of the sciences. And we might extend that view to cover more *recherché* items of ontology such as Armstrong's N or, more generally, an intrinsic causal relation or a "nexus" between causes and effects (see Menzies 1998).

A second, more direct challenge to Hume simply rejects his claim that necessity is not observable. Elizabeth Anscombe, for example, famously says: “Hume confidently challenges us to ‘produce some instance, wherein the efficacy is plainly discoverable to the mind, and its operations obvious to our consciousness or sensation’ [Hume 1739–40, 157–8]. Nothing is easier: is cutting, is drinking, is purring not ‘efficacy’?” (1971, 93). My own view here is that facts about our experience (e.g. the fact that have visual experience as of one billiard ball making another one move) simply do not settle the question about whether *what* we are observing is the kind of “nexus” between causes and effects to which Hume was so hostile (see Beebe 2003 and 2009).

Whatever the truth of the matter concerning the correct interpretation of Hume, and whether his argument for the unintelligibility of a mind-independent necessary connection between causes and effects is any good, it is certainly true that Hume has had an enormous influence on the shape of the contemporary debate about the nature, and our understanding, of causation.

4 Some Theories of Causation

This section provides an admittedly brief and partial survey of some recent theories of causation. Some theories have already been mentioned: the naïve regularity theory, the counterfactual theory, Price’s perspectivalism, Williamson’s epistemic view, and Heathcote and Armstrong’s overtly anti-Humean position. Of these, only the counterfactual theory is discussed further below. Added to the mix in this section are “process” and “mechanistic” positions and probabilistic theories of causation.

4.1 Counterfactual Theories of Causation

The fundamental insight behind counterfactual theories of causation is the very simple thought that effects *depend* upon causes in some way; and counterfactual theories cash out the notion of dependence in terms of *counterfactual* dependence. Suppose I strike a match (*c*), thereby causing it to light (*e*). Then—with some background assumptions in place, of which more later—if, contrary to what in fact happened, I hadn’t struck the match, it wouldn’t have lit. In other words, *e* *counterfactually depends on* *c*. By contrast, I step in the bath (*d*) and the neighbor rings the doorbell (*f*). If I hadn’t stepped into the bath, the neighbor would still have rung the doorbell: *f* does not counterfactually depend on *d*. This suggests a straightforward analysis of causation, as follows: *c* causes *e* if and only if had *c* not occurred, *e* would not have occurred either.

Unfortunately, things aren’t so straightforward. First, we have now replaced one mystery with another: without an account of how facts about

counterfactual dependence are determined, it's not clear that we have made very much progress. (I return to this issue below.) Second, the account just proposed cannot be right because it is subject to obvious counter-examples. In particular, it fails in cases of *pre-emption*, where there is a back-up cause waiting in the wings to cause the effect, should the actual cause fail to occur. Imagine two assassins, A1 and A2, both trying to shoot and kill the President—one from the grassy knoll, and the other from the book depository window. A2, up in the book depository, has taken aim is about to shoot when out of the corner of her eye she sees A1, out there on the grassy knoll, take aim and fire. Knowing that A1 is a crack shot (and she is right: A1 shoots and kills the President), A2 lowers her gun and makes her escape. A1's shot (*c1*) caused the President's death (*e*). But if *c1* hadn't happened, A2 would have fired instead (*c2*), and *e* would still have happened. In other words, if *c1* hadn't happened, *e* would still have happened. But that's inconsistent with the simple analysis proposed above, since clearly *c1* was a cause of *e*. Problem.

Lewis's (1973a) solution to the problem is to hold that causal *dependence* is a matter of counterfactual dependence, and that causation is a matter of a *chain* of causal dependence relations. Take some event that is on the causal path from A1's shot to the politician's death: A1's bullet piercing his heart, say. (Call this *d*.) If *c1* hadn't happened, *d* wouldn't have happened: if A1 hadn't fired her gun, her bullet never would have got there. So *d* causally depends on *c1*. And *e*, in turn, causally depends on *d*, since if the bullet hadn't pierced his heart (let's suppose), he wouldn't have been killed—our back-up assassin having given up by this point and begun to make her escape. So there is a chain of causal dependence from *c1* to *e* via the intermediate event *d*; hence *c1* is a cause of *e*, which is the right answer.

So far, so good. But there is a host of other problems lurking. One is the problem of "late pre-emption." Imagine instead that assassin A2 has been given strict instructions not to desist unless and until she sees the President well and truly dead. In that scenario, she's still on the scene with her gun trained on him at every point in between A1 firing and the President dying. So there is no event *f* in the causal chain between *c1* and *e* such that *e* counterfactually depends on *f*. For example, if A1's bullet hadn't pierced his heart (and so, we may suppose, had not fatally injured him), then the President would have died anyway because A2 would still have shot him. (Late pre-emption is so-called because the event that stops the back-up cause from kicking in is the effect itself—in this case, the President's death—whereas in "early" cases some much earlier event does the job; for example in the first case above it is A1's taking aim that stops A2 from firing.)

Another problem is that of "trumping pre-emption" (Schaffer 2000b). Imagine that there are wizards who can cast spells, and that those spells can work at a spatial and temporal distance with no intermediate states or events

that “join up” the casting of the spell with its end result. Merlin casts a spell at midnight to turn the prince into a frog, and Morgana does likewise just afterward. Imagine further that it is built into the laws of nature that if two spells are cast with the same intended effect, the later spell always cancels the first one, so that it is the second spell that is effective. So it is Morgana’s spell (*c*), and not Merlin’s, that causes the prince to turn into a frog (*e*); but—since this is a case of pre-emption—there is no counterfactual dependence of *e* on *c*: had *c* not occurred, *e* would have happened anyway, caused by Merlin’s earlier spell. By stipulation there is no intermediate event *d*, such that *d* counterfactually depends on *c* and *e* in turn counterfactually depends on *d*. So the counterfactual analysis cannot account for the fact that *c* caused *e*. Of course, there are in fact no wizards and spells; but an analysis of causation is supposed to tell us how the causal facts are determined in every *possible* situation, not merely in every *actual* situation.

Many attempts have been made to come up with a counterfactual analysis of causation that isn’t susceptible to these problems; see for example Noordhof (1999). Lewis’s own attempt (2000) involves appealing to the notion of “influence.” Roughly, the idea is that rather than conceiving causation as a matter of *whether* a given effect would have occurred had the cause not occurred, but rather a matter of the extent to which the effect would have been different had various “alterations” of the cause occurred. For example, in our “late” assassin case, while it’s true that the President would still have died had A1 not fired, the time of his death is still sensitive to the precise time at which A1 fired: had A1 fired a little earlier or later, the President’s death would correspondingly have occurred a little earlier or later. So A1’s firing has enough “influence” over the President’s death for the former to count as a cause of the latter.

4.2 Counterfactual Theories and Humeanism

In Section 3, I described counterfactual theories of causation as “Humean.” Why is that? Well, notice that I have not yet said anything about what determines the truth of the counterfactuals (“had I not struck the match, it wouldn’t have lit,” and so on) that in turn are supposed to determine the truth of causal claims. And it is in Lewis’s analysis of counterfactuals that we discover the Humean credentials of counterfactual theories of causation.

Lewis provides a *possible world analysis* of counterfactuals. Think of the whole of the Universe as the “actual world.” It is the way it is—but there are many ways it *might* have been. For example, the laws of nature might have been different. Or the laws of nature might have been the same but the Universe might have started out with different initial conditions, and so even with the same laws—things might have panned out very differently.

Think of all of the ways the world *might* have been as ways some merely *possible* world really is.

Now, some possible worlds are more similar, or “closer” (in a metaphorical sense), to the actual world than others. A possible world that is exactly similar to the actual world for the first 6 billion years of its history but then starts to diverge in similarity is much closer to the actual world than is one that started out radically different to the actual world and remains so throughout its history. More generally, Lewis (1979) holds that the closeness of one possible world to another is determined by two things: sameness of “matters of particular fact” (roughly: what happens) across regions of spacetime (a bigger region of perfect match makes for greater similarity), and sameness of the laws of nature. And Lewis’s theory of laws of nature is itself a sophisticated “regularity theory” of laws, namely a “best system” account: “a contingent generalization is a *law of nature* if and only if it appears as a theorem (or axiom) in each of the true deductive systems that achieves a best combination of simplicity and strength” (1973b, 73). To put it more crudely, for Lewis the laws of nature will turn out to be subset of the regularities: they will be those regularities that are the most powerful for making predictions.

With this account of closeness in place, Lewis’s account of counterfactuals runs as follows. Let “ $\sim O(c) \square \rightarrow \sim O(e)$ ” mean: “if event c had not occurred, event e would not have occurred.” To evaluate the truth of this counterfactual (we are assuming that both c and e did occur in the actual world), we consider the *closest* possible world w at which c does not occur—that is, the closest world at which “ $\sim O(c)$ ” is true—and see whether or not “ $\sim O(e)$ ” is true at w . If it is, then the counterfactual is true; if it isn’t, the counterfactual is false. So for example suppose our counterfactual is “if I hadn’t struck the match at t , it wouldn’t have lit.” What is the closest possible world at which I don’t strike the match (w) like? Lewis’s answer is: it is a world that perfectly matches the actual world throughout the whole of spacetime until just prior to t (until $t-1$, let’s say), and which has exactly the same laws as the actual world *except* that at $t-1$ a “small miracle” occurs: there is a minor violation of the actual world’s laws—just enough of a violation to accommodate my failing to strike the match (since, we may suppose, in the actual world the laws determine that I *do* strike the match). Thereafter, w continues to evolve according to the same laws as those that hold in the actual world. What happens after $t-1$ at w will of course start to diverge—and increasingly so—from what happens at the actual world. In particular, the match—unstruck at w —will not light, since, given the sameness of w ’s laws and the actual laws (with the exception of the miracle required to stop me striking the match), unstruck matches do not spontaneously light of their own accord at w any more than they do at the actual world.

With all this in place, we are in a position to see why Lewis’s counterfactual analysis of causation counts as a “Humean” theory. Counterfactuals are to be

analyzed in terms of what happens at the closest possible worlds, and closeness of worlds is in turn determined by (a) similarity in what happens and (b) sameness of the laws. But the laws themselves are merely a subset of the regularities. Thus, ultimately, the truth of “*c* caused *e*” is determined by the overall pattern of what happens at some close possible world, and not by the existence of any causal “nexus” between *c* and *e* in the actual world. Of course, Lewis’s theory of laws is not obligatory: for example one might, as we have already seen, hold that laws of nature are a matter of the obtaining of a necessitation relation between universals. However, it is not at all clear that such an “anti-Humean” account of laws is really compatible with Lewis’s analysis of counterfactuals. And even if it is, it remains true that Lewis’s *own* conception of the counterfactual analysis of causation is intended to be broadly Humean one, given that closeness of worlds is, for Lewis, determined solely by “Humean” facts.

4.3 Processes and Mechanisms

Worries that Lewis’s revised counterfactual analysis (2000) simply falls foul of a new set of counterexamples have been well voiced in the literature (see e.g. Schaffer 2001). A more general question to ask, however, is whether the counterfactual analysis really gets to the heart of the nature of the causal relation. One can think of Lewis’s revised account as a kind of “black box” account: we “wiggle” the input (altering the time or manner of assassin A1’s shot, say) and see what, if anything, happens to the output (the time or manner of the President’s death). But isn’t what’s really important, as far as causation is concerned, what’s going on in between the two? After all, it’s just *obvious* that A1’s shot causes the death, because we can easily trace the causal path from one to the other via the expulsion of the bullet from the gun, its movement toward the victim, its entering the victim’s heart, and so on. If the counterfactual analysis has such difficulty accounting for the causal relation between the shot and the death, perhaps that shows us that we are looking in the wrong place for an analysis of causation. Perhaps we should instead be looking inside the black box, as it were, to trace out the very obvious *process* or *mechanism* that connects the shot with the death.

This kind of approach has been taken by a number of philosophers. Perhaps the most well worked-out “process” view is that of Phil Dowe (2000), building on earlier work by Hans Reichenbach (1928), Jerrold Aronson (1971), David Fair (1979), and Wesley Salmon (1984). Reichenbach is concerned with the difference between real and “pseudo-” processes. For example, as I take a walk on a sunny day, my movement along the street constitutes a genuine causal process, with later stages causally depending on earlier stages. By contrast, the movement of my shadow is a mere pseudo-process. The later position

of my shadow is not caused by its earlier position; rather, each is independently caused by my own position. How do we account for this difference? Reichenbach's answer (1958) is that only real processes can "transmit a mark": intuitively, if one were to modify my walk by means of a single intervention—say by me picking up a large box en route—that "mark" would be transmitted to later stages of the process (so long as I don't put the box down again). By contrast, if one directly modified the *shadow* at any point (e.g. as I walk past a bus stop, the shadow acquires a bus-stop-shaped modification), that mark would *not* be transmitted to later stages: the bus-stop-shaped modification goes away again as soon as I've passed by.

Salmon (1984) refines Reichenbach's criterion for distinguishing real from pseudo-processes; however, his view has also been subjected to apparently decisive objections (see Dowe 2008, §6 for a summary). Dowe's "Conserved Quantity Theory" of causation (2000) is a version of the process theory that aims to overcome these worries. Roughly, the idea is that what is distinctive of causal processes is not that they transmit a *mark*, but that they transmit a *conserved quantity*, where a conserved quantity is any quantity identified by the laws of nature as universally conserved—according to our current best science, these would be charge, linear momentum, and mass-energy. So for example the movement of a billiard ball across the table involves the transmission of mass-energy and linear momentum from earlier to later stages, whereas a shadow is not the kind of object that can so much as possess a conserved quantity, and hence no conserved quantity can be transmitted from earlier to later stages of the shadow's movement.

One worry about the Conserved Quantity Theory (see Dowe 2000, ch.1 and Lewis 2004) is that at best it only tells us what causation *actually* consists in; it does not tell us anything about *why* the transmission of conserved quantities is what is picked out by our concept of causation. Analogy: suppose that all mental states—pain, for example—are in fact brain states (the firing of C-fibers, say). If you were an alien who knew nothing about pain, and were told that the concept *pain* picks out the firing of C-fibers, you would still be none the wiser about why *these* particular brain states are picked out by the concept of pain: you'd know which brain states constitute pain states, but you wouldn't know what it *is* to be in pain. In order to know *that*, you'd need to have something like a conceptual analysis of pain: you'd need to realize that we give the name "pain" to whatever brain states are (say) typically caused by bodily damage of some kind (being pricked with a pin or stubbing one's toe or putting one's hand in a flame, etc.) and which in turn typically cause certain kinds of behavior (removing oneself from the source of damage as quickly as possible, saying "ouch!," etc.). And the firing of C-fibers (we're imagining) is what is picked out by the concept *pain* precisely because it is the state that has just those typical causes and effects: it is the physical state that "plays the pain

role,” as it is sometimes put. (This kind of account of mental states is known as “functionalism”; see Levin 2010.)

The Conserved Quantity Theory tells us what causation is, in the same sense that neuroscience might tell us that that pain is the firing of C-fibers. It tells us nothing, however, about why the transmission of conserved quantities “plays the causation role”: it doesn’t tell us what it *is* for one thing to cause another. Peter Menzies (1996) has attempted to plug this gap by specifying the role that the transmission of conserved quantities plays: he proposes a “folk theory” of causation (including “platitudes” such as that causation is a relation between events, and that causes typically raise the probability of their effects) and suggests that the transmission of conserved quantities is that feature of the world (analogous to the firing of C-fibers in the case of pain) that in fact satisfies those platitudes.

More recently, some philosophers have offered “mechanistic” rather than process-based accounts of causation (Glennan 1996; Machamer, Darden, and Craver 2000). Machamer et al. note that “terms like ‘cause’ and ‘interact’ are abstract terms that need to be specified with a type of activity and are often so specified in typical scientific discourse. Anscombe [1971] . . . noted that the word ‘cause’ itself is highly general and only becomes meaningful when filled out by other, more specific, causal verbs, e.g., scrape, push, dry, eat, burn, knock over. An entity acts as a cause when it engages in a productive activity” (Machamer et al. 2000, 6). So the general idea is that the concept *cause* is merely an abstraction from the many different kinds of specific mechanism that we find in the world, and that are investigated by the sciences. So to uncover the nature of such a mechanism just *is* to discover the nature of one specific kind of causal relation. The mechanistic view can thus be seen as an attempt to explain Russell’s observation that the word “cause” is not to be found in the advanced sciences: such sciences have no need for such an abstract term when they can provide the nuts-and-bolts detail concerning the specific mechanisms they are concerned with.

One way to motivate a mechanistic account, as opposed to the Conserved Quantity Theory, is to note that the latter seems apt to capture causal processes visible from the perspective of physics, but is less obviously appropriate for the kinds of process studied by other sciences, such as molecular biology or the medical sciences. We need not dispute whether the transmission of conserved quantities is involved in interactions in, say, molecular biology; the point is merely that the mechanisms by which causal influence is transmitted are identified independently of consideration of conserved quantities. To use a more mundane example, you don’t need to know anything about fundamental physics to be able to identify the mechanism by which my letter posted in Manchester arrives at its intended destination in London (see Williamson 2011 for a survey of mechanistic accounts).

4.4 Probabilistic Theories

An approach to causation that appears, at least at first sight, to be very different to those described above takes its starting point from the thought that our causal talk and thought encompasses not merely causal claims about *particular* events or happenings, but also *general* causal claims: we say that smoking causes cancer, that eating sugary foods causes tooth decay, that lack of economic growth causes unemployment, and so on. From an epistemological point of view, such claims are rarely arrived at or justified by generalizing over specific cases of “particular” causation. For example, the process by which smoking causes lung cancer in any particular case is still not especially well understood, and similarly it is impossible to answer the question whether or not any particular smoker who gets lung cancer *would* have got it had they not smoked with any great degree of confidence. And yet is virtually universally agreed that smoking causes lung cancer *in general*.

How is this possible? The answer, of course, is that such general causal claims are arrived at on the basis of statistical methods: very roughly, the incidence of lung cancer among smokers is very much higher than it is among nonsmokers, and it is this general increase in the probability of lung cancer among the smoking population that justifies our belief that smoking causes lung cancer. In other words, where S = smoking and C = lung cancer, $\Pr(C/S) > \Pr(C/\sim S)$; and this is at least *prima facie* evidence that S is a cause of C .

Probabilistic theories of causation seek to analyze (or, more weakly, to shed some light on) the notion of “general” or “type-level” or “population-level” causation by appealing to this kind of statistical fact (e.g. Suppes 1970, Eells 1990; see Hitchcock 2011 for an overview). A first pass at a theory might be: A causes B (in some specified population: adult males, or UK residents, or . . .) if and only if $\Pr(B/A) > \Pr(B/\sim A)$ —that is, if and only if A raises the probability of B in that population. Unfortunately, however, this crude analysis fails in both directions. First, there can be *spurious* statistical correlations. For example, where E = a high level of educational attainment and P = attending a private school, $\Pr(E/P) > \Pr(E/\sim P)$. However, it turns out (this may be an urban myth, but it certainly *could* be true, so let’s suppose it is) that the reason for this correlation is that children’s educational attainment tends to match that of their parents and parents with a high level of attainment (A) are more likely to send their children to private schools. In other words, A is a *common cause* of E and P , and there is in fact no direct causal relation between E and P : going to a private school does not *cause* high educational attainment.

In common cause cases, if we “hold fixed” the common cause—in this case A —and consider the statistical correlation between our two effects (E and P) separately in the presence and absence of A , the correlation between E and P disappears: $\Pr(E/P \ \& \ A) = \Pr(E/\sim P \ \& \ A)$, and $\Pr(E/P \ \& \ \sim A) = \Pr(E/\sim P \ \& \ \sim A)$. So this new correlation lines up with the fact that P is not a cause of A : in our

new “reference classes” or “background contexts” (considering children with parents who have a high level of attainment and those without separately), P does not increase the probability of E . We cannot, however, simply amend our analysis of causation by stipulating that causes raise the probability of their effects in reference classes where all other *causes* of the effect are held fixed, since this would be blatantly circular. One suggested remedy (Eells 1991) is simply to hold *all* other prior factors fixed, irrespective of whether or not they are causes of the effect under consideration. One consequence of such an analysis, however, would be to divorce the analysis of “general” causation from its epistemology: in practice we never come across wholly homogeneous reference classes and so are not in a position to come to know what the statistical correlations in those reference classes are (see Dupré 1984, 1990; Eells 1987).

A related approach to general causation (see e.g. Pearl 2000; Woodward 2003), which sometimes goes under the heading of “causal modelling,” or “manipulability” or “interventionist” theories, effectively gives up on the project of providing a full-blown noncircular *analysis* of causation but nonetheless seeks to provide an illuminating account of how we can infer the causal structure of a given kind of situation from statistical evidence. The central question here is: what kinds of probabilistic relationship between a given range of variables (ranging over, say, different kinds of school, different levels of parental attainment, average income in the locality of the school, etc. if we are interested in the causes of levels of educational attainment) need to be in place in order for us to be able to infer what the *causal* relationship is between those variables?

Probabilistic theories generally can be thought of as broadly Humean theories, because they seek to analyze general causation in terms of statistical regularities. However, interventionist theories in particular fail to provide a full-blown analysis of causation because they rely on the notion of a “manipulation” or an “intervention”—and this is itself a causal notion. Here, very roughly, is why. Take our educational attainment example again. Suppose we just looked at the statistical relationships between three variables: average income in the locality of the school, type of school (state or private, let’s say), and average level of qualification of the teachers. If we did this, we might well find a statistical correlation between type of school and educational attainment. But—as we’ve already seen—this might not be a *causal* correlation: both might be effects of parental attainment, a variable that we have not included. In practice, of course, we cannot include *all* the variables that are relevant to attainment—it’s just far too complicated. So how do we ensure, with only some of the relevant variables represented in our model, that the statistical and causal correlations line up? The answer is that we consider only the statistical correlations that arise from *interventions*, where we “fix” or “set” the value of a given variable in a way that breaks any connection with prior causes

that are outside the system. In our toy example, imagine that we don't simply consider the (spurious, in fact) correlation between going to private or state school on the one hand and higher or lower educational attainment on the other. Instead, what we do is take a bunch of randomly selected 11-year-olds and send them off to private school, and send a second randomly selected bunch of 11-year-olds off to state school, and compare their attainment levels five years later. Random selection will (unless we are very unlucky) break the correlation between parental attainment and kind of school attended, since we didn't pay any attention to parental attainment when deciding which children to send to which kind of school. And so we would expect there to be no correlation between kind of school attended and the level of attainment. In effect, what we have done here is "intervened" on the variable *kind of school attended* in such a way as to screen off the effects of the usual causes of the value of that variable; and if we do *that*, the statistical correlations will line up with the causal correlations. Of course, in practice we cannot simply randomly select children and pack them off to one kind of school or another. However, this kind of practice is exactly what happens in standard randomized trials. When patients are randomly assigned to either trying out a new drug or else staying with the old one, say (without being told which group they belong to), the aim is precisely to control for other potential causes: if patients (or their doctors) make the choice themselves, spurious correlations may well emerge, just as they do if parents make the choice about which school to send their children to.

The notion of an "intervention" is, however, a causal notion, since an intervention on *X* is *defined* in terms of its breaking the causal connection between *X* and its causes. Thus this kind of probabilistic theory fails to deliver a full-blown analysis of causation. Whether or not this is a problem depends, of course, on what one's aims are in providing a theory of causation in the first place; if one's aim is to find a full-blown analysis, then an account such as Woodward's interventionist account will not, just by itself, fit the bill. On the other hand, if one's aim is to gain a practically workable theory of how to extract causal information from complex statistical relationships, it fares considerably better than other theories of causation (see Woodward 2008 for a general discussion of interventionist/manipulability accounts of causation).

5 Further Issues

Section 4 was a brief and incomplete survey of some of the most widely discussed theories of causation of recent years. In this final section, I discuss some more general issues surrounding causation that bear on the question of which theory (or theories) constitute our best account of the nature of causation.

5.1 What Does Causation Relate?

I have been talking so far of causation (or at least “particular” or “token” causation, as opposed to general causation) as though it relates *events*. But what is an event—and are events the only candidates for the relata of the causal relation? Various accounts of the nature of events exist (see, for example, Davidson 1967; Kim 1973; Lewis 1986b); I shall here briefly discuss just Lewis’s account. For Lewis (1986b), an event is a region of spacetime that has certain of its features essentially and some only contingently. Or rather, the same region of spacetime will have many Lewisian events occurring in it, all with different essential and accidental features. For example, in the region I am currently occupying, several events are currently occurring, one of which is essentially my drinking coffee, another is essentially my drinking coffee slowly, and another is essentially my drinking coffee out of a mug.

The more essential features an event has, the more “fragile” it is (so my (essentially) drinking coffee slowly (*e*) is more fragile than my drinking coffee (*f*))—in other words, the fewer possible worlds it occurs in. A possible world that is more or less like the actual world but where I am drinking my coffee quickly is a world where *f* occurs but *e* does not. The distinction between more and less fragile events is needed in order to make the counterfactual analysis of causation get the causal facts right. The reason I’m drinking my coffee slowly is that it is very hot—so its being hot is a cause of my drinking it slowly, but it is not a cause of my drinking it *simpliciter*. Likewise, my drinking the coffee slowly counterfactually depends on its being hot (if it had not been so hot, I would have been drinking it quickly), while my drinking the coffee *simpliciter* does not (if it had not been so hot, I still would have been drinking it).

Lewis imposes some restrictions on what can count as an event. In particular, events must not have “overly extrinsic” essences. Again, this is to ensure that the counterfactuals line up with the causal facts: my buying coffee just now (*c*) is a cause of my now drinking it, but it is not, intuitively, a cause of my-drinking-my-coffee-or-the-sun-exploding. If the latter were allowed to count as an event, then the counterfactual analysis would get the wrong answer, since it is true that had I not bought coffee just now, I would not now be drinking it, and nor would the sun have exploded. But my-drinking-my-coffee-or-the-sun-exploding would be an event whose essence is overly extrinsic; hence it is not really an event at all according to Lewis, and hence, since causation is a matter of counterfactual dependence between *events*, the counterfactual dependence of my-drinking-my-coffee-or-the-sun-exploding on *c* doesn’t constitute a counterexample to the counterfactual analysis.

Others (see, for example, Bennett 1988; Mellor 1995) argue that the most basic form of causal claim is “*E* because *C*,” where “*C*” and “*E*” state facts and “because” is a sentential connective. So there can be causal claims that are true but not in virtue of the obtaining of a causal *relation*. One major reason why

one might hold this view is to accommodate the (alleged) phenomenon of *causation by absence*. For example, it would seem that my failure to set my alarm clock last night caused me to sleep until 9 this morning. But my failure to set my alarm clock last night does not seem to be a candidate for an *event*—either intuitively or according to Lewis’s account of events, since to characterize an “event” as *failing* to do something is to characterize it in extrinsic terms (in terms of what is *not* going on in a given spatiotemporal region, rather than in terms of what *is* going on). By contrast, it is uncontroversially a *fact* that I failed to set the alarm clock, and so it can count uncontroversially as a causal truth that I slept until nine *because* I failed to set my alarm clock. I say more about causation by absence below.

Some authors have argued that *objects* themselves can be causes and effects. For example, one might claim that the bomb itself caused the explosion, and not merely the various events that involved the bomb in some way (my manufacturing and placing the bomb, lighting the fuse, and so on). In particular, some authors have claimed that *agents* (though not objects, such as bombs, more generally) are—or at least *could* be—causes of their own actions, in a way that does not simply reduce to their mental states (which we might think of as a species of event) causing their actions (see O’Connor 1995). Others hold that a unified account of causal relata—that is, an account that recognizes just one ontological category as a suitable relatum for causation—is preferable (Menzies 1989; see Schaffer 2008, §1 for an overview of the debate about causal relata).

5.2 Causal Relevance

When one event causes another, only some of its features—or some of the properties of the objects involved in the event—will be causally relevant to the effect. For example, suppose I sink the black ball while wearing a blue jumper and humming to myself. The angle at which I hit the cue ball and the amount of force the cue stick imparted to it are both relevant to the black going in the pocket (*e*), while neither the color of my jumper nor my humming are relevant. To put matters counterfactually, *e* would still have occurred had I not been wearing a blue jumper, or had I not been humming to myself. By contrast, it wouldn’t have occurred if I had hit the cue ball at a (very) different angle or much more gently. How should we account for the causal relevance or irrelevance of properties?

Lewis’s answer is in effect already built into his distinction between the essential and accidental features of events. *e* would still have occurred had the event that is essentially my hitting the cue ball while humming (*c1*) not occurred, since (arguably) the closest possible world where *c1* does not occur is a world where I hit the cue ball just as I did but without humming at the

same time; while e would *not* still have occurred had the event that is essentially my hitting the cue ball at the angle I did (c_2) not occurred, since the closest possible world where c_2 does not occur is a world where I hit the cue ball at a different angle and the black therefore misses the pocket. Lewis's revised (2000) account of causation complicates things rather, however; and we might also wonder whether his theory of events really gets the right answers all of the time. For example, is the closest possible world where c_3 , my hitting the cue ball while wearing a blue jumper, a possible world where I'm not wearing the blue jumper (in which case e doesn't counterfactually depend on c_3), or is it a world where I don't hit the ball at all (in which case e counterfactually depend on c_3)?

Others have argued that consideration of causal relevance reveals that causation is a *contrastive* relation: rather than simply a binary relation between two events c and e , c causes e *relative* to a given contrast class (see, for example, Maslen 2004; Schaffer 2005). Consider my sinking of the black again. Did my hitting the cue ball with force f (c) cause the black to go in the pocket (e)? The contrastivist says that the answer is: it depends what the contrast case is. If I'd hit it with just a tiny bit less force (f_1 , let's say), then I still would have sunk the black; on the other hand, if I'd hit it with a *lot* less force (f_2), the black wouldn't have made it all the way to the pocket. So relative to the contrast case f_1 , c was not a cause of e ; relative to f_2 , c *was* a cause of e .

One issue where the issue of causal relevance is crucially important is in the "causal exclusion problem." Many properties we ascribe to objects are "multiply realisable," or stand in "determinable" relations to "determinate" properties. For example, there are many ways for an object to be *red* (redness is a determinable property): it can be scarlet, maroon, pillar-box red, and so on (think of these as the determinate properties). Similarly, suppose we adopt a functionalist account of mental properties (see Section 4.3 above). Then it might well be that mental properties are multiply realizable; for example it might be that many different brain states, aside from the firing of C-fibers, can play the pain role, or that many different brain states can equally well realize the property of believing that it is raining. In fact, the vast majority of properties we normally ascribe to objects are probably determinable or multiply realizable properties: there are many ways to be square, or cold, or wet, or whatever.

This raises what is sometimes known as the "exclusion problem"—a generalized version of the problem of mental causation (see Robb and Heil 2009), which in effect is the exclusion problem as applied to the particular case of mental states. Very roughly, the problem is how to account for the causal relevance of mental properties—or of determinable or functional properties more generally. Suppose (as seems plausible) that in principle there exists a complete causal explanation of any given event that just appeals to realizer or

determinate properties. For example, if we knew enough neuroscience (say), we would be able to fully explain why Jane just pressed the button on the vending machine just by appealing to neurological or, more broadly, physical properties of Jane and her environment: her brain was in a certain state, which resulted in certain signals being sent to her muscles, which caused her finger to apply pressure to the button. But we also want to say that Jane's *wanting a cold drink*—the *mental* state she was in—was causally relevant to her pressing the button. But how can that be, when in principle we don't need to appeal to her mental state at all in giving a complete causal explanation of her pressing the button? To put it another way, the causal relevance of Jane's *neurological* state seems to exclude the possibility that her *mental* state is *also* causally relevant. Of course, in general more than one property is relevant to any given effect; both the force and the angle of my snooker shot were causally relevant to sinking the black, for example. But in this kind of case we would not have a *complete* causal explanation if we omitted one of those properties: I can't *fully* explain why the black went in the pocket without mentioning *both* the force and the angle of the shot. In the case of Jane's pressing the button, however, it seems we can make do with just the physical property; so it's unclear that we're entitled to think that mental properties (and determinable and functional properties more generally) do any causal work.

This is worrying, since in the absence of a good response to the exclusion problem we are saddled with the conclusion that determinable and functional properties are merely epiphenomenal. Various attempts have been made to resolve the exclusion problem; see for example Yablo (1992) and Bennett (2003).

5.3 One Concept of Causation or Two?

Let's return to causation by absence, mentioned above. This is an instance of what Jonathan Schaffer (2000a) calls "causation by disconnection": putative cases of causation that involve no *process* connecting the cause to the effect. Another kind of (possible) causation by disconnection is the spell-casting case described earlier. Still another is "double prevention." Suppose Jane sees the assassin aim his gun at the President, wrestles him to the ground and deprives him of the gun (*c*), thereby preventing him from shooting, which would in turn have prevented the President from arriving unharmed at his intended destination (*e*). Is *c* a cause of *e*? Certainly *e* counterfactually depends on *c*—if Jane hadn't stopped the assassin from shooting, the President would not have arrived unharmed at his destination. And the case has many other standard hallmarks of causation: for example, Jane wanted (let's suppose) to save the President's life, and her action was an excellent means for realizing that end. On the other hand, there is no causal process that connects the two events:

here is Jane wrestling the assassin to the ground, and there is the President, some distance away, oblivious to what is going on on the grassy knoll.

When we consider the kinds of theory described in Section 4 above, they fall into three different camps when it comes to causation by disconnection. Process and mechanistic theories do not recognize causation by disconnection: for example there is no transmission of any conserved quantity from *c* to *e* in the above case, and nor is there any mechanism connecting the two (for this reason Dowe argues that “causation” by disconnection is not really causation at all, but what he calls “*quasi*-causation” (Dowe 2001)). Probabilistic theories, by contrast, have no problem with causation by disconnection, because all that is important for causation is the statistical relationship between causes and effects—and such a relationship can clearly exist in the absence of any kind of connecting mechanism. (Wrestling the gunman to the ground significantly increases the probability that the intended victim will survive, for example.) When it comes to counterfactual theories of causation, one might go either way. On the one hand, counterfactual dependence just by itself does not require any kind of process connecting cause and effect—as we just saw in the Jane-and-the-assassin case. On the other hand, if we retain the idea that causes and effects must be Lewisian events, then at least some alleged cases of causation by disconnection—in particular, cases where the cause or the effect is an absence—will turn out not to be cases of causation at all, since, as we saw earlier, absences are not Lewisian events. So we might retain our commitment to the idea that causation is a relation between Lewisian events and attempt to explain away alleged cases of causation by absence (see Beebe 2004), or else—and this is Lewis’s favored approach—we might abandon the idea that causation is a relation between Lewisian events (Lewis 2004).

Ned Hall (2004) has argued that the moral we should draw is that there are in fact two distinct concepts of causation: what he calls “production” and “dependence.” Dependence is characterized in counterfactual terms and production is a relation that is transitive, intrinsic and “local”: it is constituted by the existence of a spatiotemporally continuous sequence of intermediate events. In most ordinary cases of causation the two concepts both apply, and so the existence of two distinct concepts is not immediately obvious. (When my hitting the cue ball (*c*) causes the red to go in the pocket (*e*), there is both a process leading from one to the other (production) and counterfactual dependence of *e* on *c*.) However, the two can come apart. The Jane-and-the-assassin case is one of dependence without production, and there are also cases, such as pre-emption cases (see Section 4 above), where we have production without dependence. (Hall’s proposal is discussed in e.g. Williamson 2006 and Godfrey-Smith 2009.)

Peter van Inwagen describes “cause” as a “horrible little word,” adding: “Causation is a morass in which I for one refuse to set foot. Or not unless I am

pushed" (1983, 65). Unfortunately for van Inwagen, the centrality of causation for our understanding the world *does* require us to set foot in the morass. Fortunately, many philosophers have done so. While there is no broad consensus on any of the issues discussed in this chapter, it is surely true that causation is less of a morass than it once was.

Bibliography

- Anscombe, G. E. M., 1971. *Causality and Determination: An Inaugural Lecture*. Cambridge: CUP. Reprinted in E. Sosa and M. Tooley, eds, 1993. *Causation*, Oxford: OUP.
- Armstrong, D. M., 1983. *What is a Law of Nature?* Cambridge: CUP.
- Aronson, J. L., 1971. "On the Grammar of 'Cause.'" *Synthese*, 22, pp. 414–30.
- Beebe, H., 2009. "Causation and Observation." In Beebe, Hitchcock, and Menzies 2009.
- , 2006. *Hume on Causation*. Abingdon: Routledge.
- , 2004. "Causing and Nothingness." In Collins, Hall, and Paul 2004.
- , 2003. "Seeing Causing." *Proceedings of the Aristotelian Society*, 103, pp. 257–80.
- Beebe, H., Hitchcock, C., and Menzies, P., eds, 2009. *The Oxford Handbook of Causation*. Oxford: OUP.
- Bennett, J., 1988. *Facts and Their Names*. Indianapolis: Hackett.
- Bennett, K., 2003. "Why the Exclusion Problem Seems Intractable, and how, Just maybe, to Tract It." *Nous*, 37, pp. 471–97.
- Collins, J., Hall, N., and Paul, L. A., eds, 2004. *Causation and Counterfactuals*. Cambridge, MA: MIT Press.
- Davidson, D., 1967. "Causal Relations." *The Journal of Philosophy*, 64, pp. 691–703.
- Dowe, P., 2008. "Causal Processes." In E. N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Fall edn). Available at: <http://plato.stanford.edu/archives/fall2008/entries/causation-process/>
- , 2001. "A Counterfactual Theory of Prevention and 'Causation' by Omission." *Australasian Journal of Philosophy*, 79, pp. 216–26.
- , 2000. *Physical Causation*. Cambridge: CUP.
- Dupré, J., 1990. "Probabilistic Causality: A Rejoinder to Ellery Eells." *Philosophy of Science*, 57, pp. 690–8.
- , 1984. "Probabilistic Causality Emancipated." In P. French, T. Uehling, Jr., and H. Wettstein, eds, *Midwest Studies in Philosophy IX*. Minneapolis: University of Minnesota Press, pp. 169–75.
- Eells, E., 1991. *Probabilistic Causality*. Cambridge: CUP.
- , 1987. "Probabilistic Causality: Reply to John Dupré." *Philosophy of Science*, 54, pp. 105–14.
- Fair, D., 1979. "Causation and the Flow of Energy." *Erkenntnis*, 14, pp. 219–50.
- Glennan, S. S., 1996. "Mechanisms and the Nature of Causation." *Erkenntnis*, 44, pp. 49–71.
- Godfrey-Smith, P., 2009. "Causal Pluralism." In Beebe, Hitchcock, and Menzies 2009.

- Goldman, A., 1967. "A Causal Theory of Knowing." *The Journal of Philosophy*, 64, pp. 357–72.
- Grice, H. P., 1961. "The Causal Theory of Perception." *Proceedings of the Aristotelian Society*, Supp. Vol. 35, pp. 121–53.
- Hall, N., 2004. "Two Concepts of Causation." In Collins, Hall, and Paul 2004.
- Heathcote, A. and Armstrong, D. M., 1991. "Causes and Laws." *Nous*, 25, pp. 63–73.
- Hitchcock, C., 2011. "Probabilistic Causation." In E. N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Winter edn). Available at: <http://plato.stanford.edu/archives/win2011/entries/causation-probabilistic/>
- Hume, D., 1748/1751. *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, ed. L. A. Selby-Bigge, 1975. 3rd edn, revised and ed. P. H. Nidditch, Oxford: Clarendon.
- , 1739–40. *A Treatise of Human Nature*. Edited by L. A. Selby-Bigge, 1978. 2nd edn, revised and ed. P. H. Nidditch, Oxford: Clarendon.
- Kim, J., 1973. "Causation, Nomic Subsumption, and the Concept of Event." *The Journal of Philosophy*, 70, pp. 217–36.
- Levin, J., 2010. "Functionalism." In E. N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Summer edn). Available at: <http://plato.stanford.edu/archives/sum2010/entries/functionalist/>
- Lewis, D. K., 2004. "Void and Object." In Collins, Hall, and Paul 2004.
- , 2000. "Causation as Influence." *The Journal of Philosophy*, 97, pp. 182–97. Reprinted in Collins, Hall, and Paul 2004.
- , 1986b. "Events." In Lewis 1986a.
- , 1986a. *Philosophical Papers, Vol. II*. Oxford: Blackwell.
- , 1981. "Causal Decision Theory." *Australasian Journal of Philosophy*, 59, pp. 5–30.
- , 1979. "Counterfactual Dependence and Time's Arrow." *Noûs*, 13, pp. 455–76. Reprinted in Lewis 1986a.
- , 1973b. *Counterfactuals*. Oxford: Blackwell.
- , 1973a. "Causation." *The Journal of Philosophy*, 70, pp. 556–67. Reprinted with postscripts in Lewis 1986a.
- Machamer, P., Darden, L., and Craver, C., 2000. "Thinking about Mechanisms." *Philosophy of Science*, 67, pp. 1–25.
- Mackie, J. L. 1965. "Causes and Conditions." *American Philosophical Quarterly*, 2, pp. 245–64. Reprinted in E. Sosa and M. Tooley, eds, 1993. *Causation*, Oxford: OUP.
- Maslen, C., 2004. "Causation, Contrasts, and the Nontransitivity of Causation." In Collins, Hall, and Paul 2004.
- Mellor, D. H., 1995. *The Facts of Causation*. New York: Routledge.
- Menzies, P., 1998. "How Justified are the Humean Doubts about Intrinsic Causal Links?" *Communication and Cognition*, 31, pp. 339–64.
- , 1996. "Probabilistic Causation and the Pre-emption Problem." *Mind*, 105, pp. 85–117.
- , 1989. "A Unified Account of Causal Relations." *Australasian Journal of Philosophy*, 67, pp. 59–83.
- Menzies, P. and Price, H. 1993. "Causation as a Secondary Quality." *British Journal for the Philosophy of Science*, 44, pp. 187–203.

- Noordhof, P., 1999. "Probabilistic Causation, Preemption and Counterfactuals." *Mind*, 108, pp. 95–125.
- O'Connor, T., 1995. "Agent Causation." In T. O'Connor, ed., *Agents, Causes, and Events: Essays on Indeterminism and Free Will*. New York: OUP, pp. 173–200.
- Pearl, J., 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: CUP.
- Price, H., 2007. "Causal Perspectivalism." In H. Price and R. Corry, eds, *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford: OUP, pp. 250–92.
- Quine, W. V. O., 1963. "Reference and Modality." In *From a Logical Point of View*. New York: Harper & Row, pp. 139–59.
- Ramsey, F. P., 1929/1990. "General Propositions and Causality." In D. H. Mellor, ed., *F. P. Ramsey: Philosophical Papers*. Cambridge: CUP, pp. 145–63.
- Reichenbach, H., 1928. *The Philosophy of Space and Time*. 1958 edn. Translated by M. Reichenbach and J. Freund. New York: Dover.
- Robb, D. and J. Heil., 2009. "Mental Causation." In E. N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Summer edn). Available at: <http://plato.stanford.edu/archives/sum2009/entries/mental-causation/>
- Russell, B., 1912–13. "On the Notion of Cause." *Proceedings of the Aristotelian Society*, 13, pp. 1–26.
- Salmon, W., 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.
- Schaffer, J., 2008. "The Metaphysics of Causation." In E. N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Fall edn). Available at: <http://plato.stanford.edu/archives/fall2008/entries/causation-metaphysics/>
- , 2005. "Contrastive Causation." *The Philosophical Review*, 114, pp. 297–328.
- , 2001. "Causation, Influence, and Effluence." *Analysis*, 61, pp. 11–19.
- , 2000b. "Trumping Pre-emption." *The Journal of Philosophy*, 97, pp. 165–81.
- , 2000a. "Causation by Disconnection." *Philosophy of Science*, 67, pp. 285–300.
- Suppes, P., 1970. *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Publishing Company.
- van Inwagen, P., 1983. *An Essay on Free Will*. Oxford: Clarendon.
- Williamson, J., 2011. "Mechanistic Theories of Causality." *Philosophy Compass*, 6, pp. 421–47.
- , 2006. "Causal Pluralism versus Epistemic Causality." *Philosophica*, 77, pp. 69–96.
- Woodward, J., 2008. "Causation and Manipulability." In E. N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Winter edn). Available at: <http://plato.stanford.edu/archives/win2008/entries/causation-mani/>
- , 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: OUP.
- Yablo, S., 1992. "Mental Causation." *The Philosophical Review*, 101, pp. 245–80.

22 Metaphysics

E. J. Lowe

In this chapter, I propose to focus on three key notions: those of *identity*, *change*, and *modality*. These notions are both intimately connected with each other and centrally important to metaphysics and ontology and their relationship with logic. By the notion of “modality,” here, I mean the family of notions that most importantly includes those of *possibility*, *necessity*, and *actuality*. When we speak of *change*, I think that we always do so in the context of something that is presumed *not* to change and thus to remain the *same*—that is, *identical*. To attribute change, moreover, presupposes that *what* the change is attributed to *can* change—that it is *possible* for it to change—in the manner that is attributed to it: and whether that is so will depend on what *kind* of thing it is that the change in question is being attributed to. Things of some kinds, we believe, can *persist* through certain types of change which, in things of other kinds, would result in their *ceasing to exist* altogether—persistence being nothing less than *identity over time*, or so-called diachronic identity. But what is the ontological basis of such possibilities and how do we have knowledge of them? These are some of the main issues that I shall be discussing in this chapter, which is appropriately divided into three parts: the first on identity, the second on change, and the third on modality. Besides trying to resolve certain important problems that these issues raise, I hope that my manner of tackling these problems will serve to show, more generally, how I think that questions of metaphysics should be approached and what makes them distinctively different from questions of other kinds, including empirical scientific questions. In that respect, at least, this chapter is intended to be a contribution to *metametaphysics*, or the methodology of metaphysics, quite as much as it is to metaphysics itself.

Part 1 Identity

1.1 Two Senses of “identity”

What is *identity*? I believe that this term has two distinct but related senses in metaphysics and logic. (a) In one sense, “identity” denotes an important *formal relation*, the identity relation, which logicians symbolize by the equality

sign, “=” In this sense, we speak of something, *a*, being, or not being, identical with something “else,” *b*. Of course, if *a* and *b* really are (numerically) identical, then *b* is *not* something “else” than *a*. I call identity in this sense a *formal* relation not just because it is a relation recognized by formal logic, but in order to distinguish it from “real” or “material” relations, such as spatial and perhaps causal relations. If *a* is identical with *b*, then *a* (and so *b*) exists, but it would be wrong to suppose that, in addition, a real *relation* exists between *a* and itself. (b) In another sense, “identity” denotes what is sometimes called by metaphysicians “individual essence.” This is the sense in use when we speak of some *thing’s* “identity,” such as the identity of *a*. The identity of *a*, in this sense, is *what a is*, or *what it is to be a*, as opposed to any other thing (see Lowe 2008b). Note that the expression “what *a* is” should not be confused with the expression “what *makes a* what it is,” which denotes the *individuator* of *a*—an entity that ought not to be conflated with *a*’s individual essence, although it will be part of *a*’s individual essence *that it has* this individuator (see Lowe 2003). Until recently, modern metaphysicians were very skeptical about this second notion of identity, but now they are beginning to change their minds. As we shall see in Part 3, the notion of individual essence is very important for understanding *modal* notions. For the purposes of the present Part, however, I shall be concentrating on “identity” in its *first* sense, that is, on identity as a formal relation.

1.2 The Logical Laws of Identity

As a formal relation, identity might seem to be rather uninteresting, because it has just a few very evident logical properties. These can be reduced to two: *being reflexive* and *being subject to Leibniz’s Law*. A relation, *R*, is reflexive just in case, for anything, *x*, which stands in that relation, *x* stands in *R* to itself. Leibniz’s Law is the principle that, for anything, *x*, and anything, *y*, if *x* is identical with *y*, then whatever is true of *x* is also true of *y*. From the fact that identity is reflexive and subject to Leibniz’s Law, it can easily be shown that identity must have two other logical properties: *being symmetrical* and *being transitive*. As a consequence, identity is an *equivalence* relation (reflexive, symmetrical, and transitive). And because of Leibniz’s Law, it is the “smallest” equivalence relation—that is to say, all of the “equivalence classes” of the relation are necessarily *single-membered*, each such member being an ordered pair whose own members are just a certain thing *and itself*, $\langle x, x \rangle$. In other words, identity is the relation in which each thing necessarily stands to itself and to no other thing. Could anything be more straightforward and less mystifying than that? And yet, as we shall see, the notion of identity as a relation generates all sorts of deep philosophical puzzles.

1.3 Criteria of Identity

The next notion that we need to introduce, in order to prepare for these puzzles, is that of a *criterion of identity*, which we owe in large measure to Frege (see Lowe 1989a, ch. 2, and Lowe 1989b). There are two different ways of thinking of identity criteria: (a) as *epistemic* or *heuristic* principles, guiding us with regard to the *evidence* we should appeal to in our judgments of identity; and (b) as *metaphysical* principles telling us what identity “consists in” (to use Locke’s phrase) for entities of various different kinds. Thus, for instance, fingerprints provide reasonably good evidence of the identity of human beings, if we seek to identify a given suspect with the perpetrator of a certain crime. But they don’t tell us what human personal identity *consists in*—a very difficult philosophical question. In what follows, I shall only be concerned with the *metaphysical* notion of an identity criterion. The general form of any such criterion is this: “If x and y are entities of kind K , then x is identical with y if and only if x and y stand in relation R_k ”—where this relation is not, of course, simply that of identity itself, since then the criterion would be trivial and circular. (Note that this is not in fact the form of “Fregean” identity criteria, which is, rather: “The K of x is identical with the K of y if and only if x and y stand in relation R_k ”—for instance, “The direction of (line) x is identical with the direction of (line) y if and only if x and y are parallel”: see Lowe 1989b.) We can call R_k the *critical* relation for entities of kind K —and the important point is that this relation may well be different for entities of different kinds.

Some philosophers, however, are skeptical about the very notion of an identity criterion. Some think that Leibniz’s Law itself provides all that could possibly be needed in the way of such a criterion, in the metaphysical sense. Others think that those who propose that entities of different kinds are governed by different identity criteria are mistakenly supposing that “identity” is differently defined for different kinds of entity, and is thus an ambiguous expression. Neither of these objections is sound. An identity criterion is not a *definition* of identity: in fact, I think that “identity” is a perfectly univocal but indefinable, because primitive, expression. What such a criterion really tells is not something about the notion of *identity*, but something about the nature of the kind K of entities to which it applies. For example, in order to grasp properly the mathematical notion of a *set*, we need to understand the criterion of identity for sets, which is just this: “If x and y are sets, then x is identical with y if and only if x and y have the same members.” The *same membership* relation is the critical relation for sets, and this tells us something about the nature of sets. Note, here, that this relation itself involves *identity*, but not the identity of *sets*—only of their members—which is why it is not trivial and circular.

1.4 The Absoluteness of Identity

One important confusion that we must dispel at this point is this. We shouldn't suppose, just because entities of different kinds are governed by different criteria of identity, that identity *itself* must be relativized to different kinds—that it is “sortally relative.” That is to say, we shouldn't suppose that entities *a* and *b* could be identical as entities of kind K_1 , but distinct as entities of kind K_2 . The notion of identity is an *absolute*, not a relative, one. Some philosophers—notably Peter Geach (1980)—have denied this, saying, for instance, that the same lump of bronze could be, at different times, two different statues. (Note that it is a consequence of this view that Leibniz's Law does not, in fact, hold unrestrictedly.) What I think we should say, following David Wiggins (1980; 2001), is just that the same lump of bronze may, at different times, *constitute* two different statues—not that it can *be* two different statues (in the sense of “be” that corresponds to the “is” of identity). However, we shall return to this issue in Part 2, so I shall not dwell any further on it here.

1.5 The Puzzle of Persistence

Philosophical puzzlement about identity generally begins when we introduce *time* into the picture—as, indeed, is illustrated by the foregoing example of the two statues. This, I think, is really because *time* is puzzling, not because *identity itself* is puzzling. However, many philosophers talk, in this context, in terms of a distinction between “synchronic” and “diachronic” identity—identity *at* a time, as opposed to identity *over time*. Now, once again, it would be a mistake to think of these as different species of *identity*: for “identity” is univocal. But why, then, is identity “over time” at all puzzling? First, let us be clear that talking about identity “over time” is just another way of talking about *persistence through time*. An object, *a*, *persists* through a period of time, from t_1 to t_2 , just in case *a*—that *very same object*—exists at *every* moment of time between t_1 and t_2 . But *how*, it may be asked, can the *very same* object exist at *different* times? This strikes some philosophers as being deeply mysterious. Perhaps this is because they think that having the same object exist at different times would be rather like having the same object existing in different *places* at the *same* time. And isn't it obvious that an object cannot be in two different places at once? Well, that depends on what exactly is meant by this. Clearly, a material object, such as an apple, cannot be *wholly* in two different places at once—all of it in the kitchen and *all* of it in the dining room, say. (Unless, perhaps, time travel is possible!—but we won't go into that now.) But it could be *partly* in the kitchen and *partly* in the dining room—for instance, if it is cut in half and each half is in one of the two rooms, or if the rooms have an interconnecting door and the apple is situated in the middle of the doorway. This line of thought thus leads some philosophers to suggest something similar concerning identity

over time: they say that the same object can exist at different times only if it has different *parts* that exist at those times. These “parts,” however, will have to be *temporal*, rather than *spatial*, parts of the apple.

1.6 Time and Physical Theory

The *doctrine of temporal parts*, as I have just presented it, seems to presuppose that time is relevantly similar to space—that time, like each of the three directions of space, is a physical *dimension* in which things can be *extended* and thus have parts. It is sometimes claimed that this view of time is unavoidable in the light of Einstein’s special and general theories of relativity, which take *spacetime* to be a unitary four-dimensional manifold. However, it seems that even Einstein himself did not adopt this view until persuaded to by the work of Minkowski—it is not a feature of his original presentation of the special theory of relativity in 1905. This is really more a question for metaphysics than for physics as such. And, in any case, Einstein’s theories *are* only theories. They may be empirically well-confirmed, but that doesn’t rule out non-standard interpretations of these theories that have different metaphysical implications (see Lowe 2002, pp. 268–70). So we cannot simply assume that the doctrine of temporal parts is made inevitable by developments in physical theory.

1.7 Primitive Persistence and Temporal Parts

It is not yet clear, then, that we *must* subscribe to the doctrine of temporal parts in order to make sense of the idea of the same object existing at different times, and thus in order to make sense of the notion of persistence through time. However, in Part 2 we shall examine a line of argument that appeals to the possibility of *change* over time in order to defend the doctrine of temporal parts—an argument focusing on the so-called problem of temporary intrinsics, which we owe to David Lewis. Meanwhile, let me try to connect the notion of a *criterion of identity* to that of persistence. If it is in the nature of a kind of entities, *K*, to persist through time, then, since persistence is just a matter of identity over time, an adequate criterion of identity for *Ks* ought to tell us what the *persistence conditions* of *Ks* are. By the “persistence conditions” of *Ks*, I just mean the conditions that are logically necessary and sufficient for an entity of kind *K* to persist through a given period of time. Now, in many cases, we have a good intuitive idea of what such persistence conditions are. Suppose, for instance, that we are talking about the following kind of entity: *an individual pile of stones*. Such a pile is composed, at any given time, of a number of different stones, all heaped together. What is required in order for the *same* pile of stones to persist through a given period of time? Just this,

surely: the *same stones* that composed that pile at the start of the period should *continue to be heaped together* throughout the period.

But notice that we can, in this way, account for the persistence of the pile only in terms of the persistence of the individual stones that compose it. This rather suggests that—on pain of an infinite regress—there must be *some* persisting things whose persistence is *primitive*, in the sense that it is not explicable in terms of the persistence of things of any other kind. These things might be, perhaps, the “fundamental particles” of physics. But what, now, about the doctrine of temporal parts? Doesn’t that maintain that we can give an account of the persistence conditions even of these fundamental particles—not in terms of the persistence of any other kind of thing, but rather in terms of relationships between their supposed temporal parts? For, of course, temporal parts—at least, *momentary* temporal parts—are not supposed to be persisting things. That is indeed so. The temporal parts theorist believes that a *reductive* account of persistence is always available, in terms of relations between *nonpersisting* things. The relations in question will be *spatiotemporal* and (perhaps) *causal*. Is that belief reasonable? Well, as I have mentioned, we shall look at an argument for it in Part 2.

1.8 Persistence and Time-order

Meanwhile, however, we might raise the following general doubt. If persistence always consists in relations between nonpersisting things that exist at different times, as the temporal parts theorist maintains, what entitles us to suppose that those nonpersisting things really *do* stand in the required relations? For instance, if *a*’s persistence from t_1 to t_2 depends on there being nonpersisting entities *b* and *c*, such that (a) *b* exists only at t_1 , (b) *c* exists only at t_2 , and (c) t_1 is *earlier* than t_2 , then what *makes it true* that t_1 is earlier than t_2 ? More generally, what makes it true that two different times, t_1 and t_2 , belong to *the same time-order*, so that either they are simultaneous or one is earlier than the other? If it can be argued that different times belong to the same time order only if something existing at one of them also exists at the other—in other words, only if something *persists* between those two times—then, clearly, it would be impossible to maintain, without circularity, that persistence always consists in relations between nonpersisting things that exist at different times.

Now, a temporal parts theorist might maintain that time-order is generated by, or constituted by, *causal* relations, not by the *persistence* of anything over time. However, that seems to invert the proper order of explanation, since causal relations can only obtain between entities existing at the same or different times *in the same time-order*. After all, if we were to suppose that two entities existed in *two quite different worlds*, each with its own space and time, then, of course, we would have to conclude that neither could causally affect

the other. In other words, causal relations of the requisite sort appear to *pre-suppose*, and so cannot *explain*, time-order. Perhaps only *persistence* can. We needn't suppose, however, that there is any *single* entity that persists throughout *all* the times in the same time-order, just that for any moment in that time-order, there is *some* entity that persists through it. Thus, the persisting entities whose collective existence secures the unity of a world's time-order, making the world *one* world in time, could be like the overlapping fibers in a rope—no single fiber needs to extend throughout the whole length of the rope, so long as *some* fiber extends through each cross-section of it (see Lowe 2009).

Part 2 Change

2.1 The Varieties of Change

There are many different kinds of change, but here I shall focus on three important species of change that can be undergone by persisting objects. These are (a) *compositional* change, or change of *parts*, (b) *qualitative* change, or change of *properties*, and (c) *substantial* change, or change with respect to *existence*. When I speak of change of “parts,” I do not, of course, mean *temporal* parts, for nothing possessing temporal parts could intelligibly be said to undergo a *change* in respect of those parts. To make this clear, let us speak, in this context, of *component* parts, where the component parts of a persisting object are always assumed to be *other persisting objects*. Thus, for example, the top and legs of a table are among its component parts. As for qualitative change, this seems straightforward enough, at least at first sight. A persisting object undergoes such change just in case it possesses different properties at different times at which it exists. For instance, if a table is colored red at one time, but colored white at a later time, then it has undergone a qualitative change, in respect of its color.

However, just as philosophers sometimes slip into the error of supposing that “synchronic” and “diachronic” identity are different kinds of *identity*, so too they sometimes make the error of supposing that, in addition to “numerical” identity, there is another kind of identity—“qualitative” identity—such that two objects that are numerically distinct may none the less be qualitatively identical, at least in some respect. Thus two different tables might be said to be qualitatively identical in respect of their color, if both were colored the same shade of red. But it should be clear that what we are really talking about here is just the *numerical* identity of *their colors*, so that we are not really talking about a different kind of identity when we speak of “qualitative” identity—just about the numerical identity of *qualities*, as opposed to that of the objects possessing them. (A further complication here, however, is that there is debate among metaphysicians as to whether, in fact, different objects

can possess numerically the same quality: those who think that qualities are universals will say “Yes,” while those who think that they are “tropes” will say “No”: but I set aside this dispute for present purposes).

Finally, then, there is *substantial* change, or *existence* change, which occurs to a persisting object when it either begins or ceases to exist.

2.2 The Paradoxes of Compositional Change

The idea of compositional change has given rise to some of the most notorious philosophical paradoxes, the best-known example being that of the ship of Theseus. We start with the observation that persisting things of certain kinds, such as ships, seem to be capable of undergoing a change of their component parts—unlike such things as a pile of stones, which apparently requires to be composed of the same individual stones throughout its existence. However, we then realize that no *limit* to the extent of such compositional change can consistently be set: if a ship can change *some* of its parts and still exist, then it can, sooner or later, change *all* of them and still exist. If we deny this, we run into a conflict with the law of the transitivity of identity. We can still insist, of course, that a ship cannot change all of its parts *at the same time*, only gradually and successively. Once we do allow this, however, we must allow that a ship could survive such a complete change of parts *without the destruction of those parts*, which could therefore be reassembled to form an exactly similar but numerically distinct ship. We now seem to have two rival candidates for identity with the original ship: the “renovated” ship, possessing all the replacement parts, and the “reconstructed” ship, possessing all the replaced parts. This is the puzzle as Thomas Hobbes presents it to us in his *De Corpore*.

One rather drastic response to the puzzle is to endorse the doctrine of *mereological essentialism*, as Roderick Chisholm (1976, ch. 3) has done. This is the view that *no* composite object can literally survive a change of any of its component parts. Rather, the result of such a change—which is consequently really a *substantial* change—is that one composite object ceases to exist and another, with some but not all of the same parts, comes into existence in its place. So what we really have is not a single, persisting composite object, but a *succession* of them, which we *mistake* for being a single thing. (But, be it noted, this account is *not* a version of the temporal parts theory of persistence, despite having a superficial similarity to it.) This approach is very much like Hume’s view of “identity” over time as being a “fiction”—although Chisholm would say this only of composite things that seem to undergo a change of parts.

It would be preferable, I think, to solve the puzzle of the ship of Theseus less drastically. My own solution, which I shall only touch on here, is to say that it is only the *renovated* ship that has any claim to identity with the original ship, because it alone can be credited with having a *continuous* existence

throughout the process of part-replacement. For it seems clear that we do not have *two distinct ships* until this process is complete, or at least nearly complete, and that at earlier stages during the process the only ship in existence is one that possesses some, but not all, of the original ship's parts (see Lowe 2002, ch. 2). Consequently, I do not think that this sort of puzzle concerning compositional change presents a serious challenge to our ordinary concepts of persistence. However, as we shall see a little later, there are other such puzzles that may be more challenging, such as the puzzle of Tibbles and Tib (or, in its ancient version, of Dion and Theon). These, however, have more to do with the problem of *substantial* change.

2.3 The Problem of Temporary Intrinsics

In the case of qualitative change, we face the problem—mentioned in Part 1—of *temporary intrinsics*, which we owe to David Lewis (1986, pp. 202–4). The question is: how can an object *a* possess, at time t_1 , a property *F*, but also possess at a later time, t_2 , another property *G* that is incompatible with *F*? How, for instance, can the same table be successively *red* and *white*, when red and white are mutually incompatible colors? Surely, one and the same thing cannot be both red *and* white, or both round *and* square? Indeed, is this not an immediate implication of Leibniz's Law which, as will be recalled from Part 2, is the principle that if *x* is identical with *y*, then whatever is true of *x* is also true of *y*? For if it is true of the table existing at the earlier time that it is red, and that table is *identical* with the table existing at the later time, must it not also be true of the latter table that it is red—and hence *not* white? Here the obvious reply is that this misconstrues Leibniz's Law. There is just one table, existing at both times, and it is true of this table both that it *was* red at the earlier time and *is* white at the later time: and *being red at the earlier time* is not incompatible with *being white at the later time*, only with *being white at the earlier time*.

So what is the problem? Well, Lewis contends that to respond in this fashion is to ignore the fact that redness and whiteness are supposed to be *intrinsic* properties of the things that have them, not *relational* properties. Instead of talking about *being red* and *being white*, we are now talking about *being red at the earlier time* and *being white at the later time*, and these are relational properties, involving *relations to times*. He concludes that the only things that can really have the intrinsic properties of *being red* and *being white* are the *temporal parts* of the table that exist at the earlier and later times respectively. Thus he regards the problem of temporary intrinsics as requiring the doctrine of temporal parts for its solution. There is, he acknowledges, another possible solution, which is to espouse the doctrine of *presentism*, by denying the existence of any time but the present time and of any thing but presently existing things. But he regards this as being too high a price to pay for solving the problem.

I, on the other hand, am strongly inclined to regard the problem as a *pseudo*-problem (to borrow Carnap's expression). When Lewis says that the first solution offered above wrongly treats the color properties of persisting objects as *relational* rather than intrinsic properties, he is trading on a false analogy. Ordinarily, we say that a property of an object is "relational" if the object needs to stand in some relation to *another object* in order to possess the property. The property of *being a father* is relational in this sense. But *times* are not just *other objects* in this sense. Indeed, it is not clear that we should *reify* times at all. It is true that our intuitive concept of a color property like redness is that it is not relational in the way that fatherhood is. But I don't think that we can infer from this that our intuitive concept of redness is that it is an "intrinsic" property in the sense that Lewis requires—that is, that it is a property that must be ascribable to an object without any reference to the time at which it is possessed by that object. This, in a way, is the point of the so-called adverbial solution to the problem of temporary intrinsics (see Lowe 1998, pp. 127–35).

The upshot of all this is that I see no good reason, based on that problem, for believing in the doctrine of temporal parts. For that matter, neither do I see any good reason to believe in presentism on the grounds that it has a solution to that problem. As I have just indicated, I think that what we are faced with here is nothing more than a pseudo-problem, not a puzzle that requires us to adopt a radically new ontology, whether that be one of temporal parts or one denying the reality of anything but the present.

2.4 Substantial Change and the Puzzles of Material Coincidence

As I explained earlier, a *substantial* change is one that occurs when something either begins or ceases to exist. This might seem, in itself, to be a completely unproblematic notion. However, in this case too philosophers have managed to think of puzzles and paradoxes. One of the best known is the problem of Tibbles and Tib (see Lowe 1989a, ch. 6). This is sometimes called the *paradox of decrease*, its converse being the *paradox of increase*. As we noted earlier, it seems intuitively correct to acknowledge that composite objects can sometimes undergo change in respect of their component parts—compositional change. Previously, in the case of the ship of Theseus, we considered only a case of compositional change *with replacement of parts*. But, equally, it seems possible that an object should simply lose a relatively minor part and continue to exist. Indeed, even when replacement occurs, there must in general be a small time-interval between the removal and the replacement of a part, through which the composite object persists.

Consider then Tibbles the cat and its tail, Tail, which we assume that Tibbles can lose without ceasing to exist. And let us call all of the rest of Tibbles, apart from his tail, "Tib." Tibbles and Tib are clearly numerically distinct objects,

possessing different properties and parts. Moreover, since Tib does not contain Tail as a part, Tib should surely be able to survive the removal of Tail from Tibbles. However, this means that when Tail is removed from Tibbles, *both Tibbles and Tib survive*. Now, though, Tibbles and Tib are *materially coincident* objects, occupying exactly the same place at the same time and being composed of exactly the same parts. This has seemed to many philosophers a repugnant and even incoherent conclusion, which they feel compelled to deny. They must therefore question one of the premises leading to it, or else the validity of the inference that has been drawn.

Some argue that it is a mistake to suppose that there ever was such an object as “Tib,” or indeed one such as “Tail”—they deny what Peter van Inwagen (1981) calls the “doctrine of arbitrary undetached parts.” Others, such as Michael Burke (1994), argue that, contrary to what was earlier assumed, Tib must cease to exist upon the removal of Tail from Tibbles. Yet others, however, maintain that we should simply accept that Tibbles and Tib can exactly coincide, since it is plausible to maintain this in other cases, such as that of a lump of bronze and the bronze statue that it constitutes—recall from Part 1 that this is what Wiggins says in this case, in opposition to Geach’s theory of relative identity. Now, I agree with Wiggins in this latter case, but don’t think that such a view can easily be extended to cover the case of Tibbles and Tib. This is because statues and lumps of bronze very plausibly have *different criteria of identity* and hence different persistence conditions, but it is not so easy to maintain this in the case of Tibbles and Tib. Moreover, if we once accept that Tibbles and Tib can be two materially coincident objects, we shall have to accept, by the same process of reasoning, that *any finite number*, no matter how large, of material objects can coincide with one another, since there are indefinitely many minor parts of Tibbles that could be removed without Tibbles ceasing to exist. For this reason, I think that there is a good deal to recommend the proposal that “Tib” was a fiction right from the start: there never was such an object.

2.5 The Bronze Statue and the Lump of Bronze

As I have just indicated, while I don’t think that Tibbles and Tib provide a genuine example of materially coincident objects, I do think this in the case of the bronze statue and the lump of bronze of which it is made. The lump doesn’t *cease to exist* when the statue is formed from it, as Burke supposes. However, some philosophers, such as Eric Olson (2001), still protest that they cannot understand what could *make* the statue and the bronze distinct objects, given that they are composed of exactly the same bronze particles arranged in exactly the same way. How, they ask, can a single arrangement of the same particles support the existence of *two different objects* composed by those particles? Are

not all the properties of any object composed by those particles determined by the properties and relations of those particles? So how can they determine one object to be a *statue* and another to be a *lump*, with different criteria of identity and hence different persistence conditions?

My reply to these philosophers is that they misunderstand the nature of an object's persistence conditions and also misunderstand the nature of material composition (see Lowe 2002, ch. 4, and Lowe 2008a). Criteria of identity and principles of composition are not empirical, physical properties of objects, like color or shape or solubility in sulfuric acid. Rather, they are a priori metaphysical principles that we understand only when we grasp the *essences* of the kinds of objects to which they apply. I shall say much more about essence in Part 3. But the basic point is that it is through understanding what a statue *is* and what a lump of bronze *is* that we see that they have different persistence conditions and different principles of composition. Moreover, these principles of composition do not make reference to the component parts of an object at only a *single* moment of its existence. Although, at such a moment, the statue and the lump may have exactly the same bronze particles as parts, the reason why those particles compose a *statue* at that time differs from the reason why they compose a *lump* at that time. They compose a *lump* then simply because *all of them* have for some time *adhered together*, but that is not the reason why they compose a *statue* then. They compose a statue then only because they *or other particles* have for some time adhered together *in a certain overall shape*, fit to represent what the sculptor intended. *Of course* one cannot deduce, from the mere fact that these particles have a certain arrangement at a certain time, that they then compose two different objects, a statue and a lump. Rather, it is a *consequence* of the different composition principles of statues and lumps that at one and the same time the very same particles, arranged in a certain way, may simultaneously compose both a statue and a lump. These are principles that we have to grasp if we are to understand what statues and lumps *are*, not principles that we could hope to infer from a knowledge of how material particles are arranged at any given time, in the way that one might hope to infer the empirical properties of statues and lumps, such as their color or electrical conductivity.

2.6 The Modal Character of Identity Criteria and Composition Principles

One important lesson that we can draw from our discussions so far is that both identity criteria and composition principles are a priori and *modal* in character. They express *necessary* truths that *we can know independently of experience*. Both types of principle are restricted to appropriate *kinds* of object. We understand the principles by understanding the *natures or essences* of the kinds in question. For instance, it is because we understand what a pile of stones *is* that

we can see that it—that very same pile—*cannot continue to exist* if one or more of the stones in it is replaced or destroyed. Similarly, it is because we understand what a statue *is* that we can see that, unlike a lump of bronze, it *cannot survive a radical alteration in its shape*. There is really nothing at all mysterious about such knowledge. It only seems mysterious to philosophers who misguidedly assume that all modal truths must somehow depend or “supervene” upon empirically ascertainable truths about the actual physical properties and relations of material objects. This, however, leads us naturally to the topic of Part 3—*modality*.

Part 3 Modality

3.1 Possible Worlds and “transworld identity”

At the end of Part 2, I remarked that criteria of identity, along with principles of composition, are *necessary* and *a priori* in character. They reveal to us something concerning the *essences or natures* of the kinds of objects to which they apply—statues, lumps, piles of stones, sets, or whatnot. But now that we have entered the domain of modal truths—truths about what is necessarily or possibly the case—we have to inquire more closely into the *grounds* of such truths and our ability to know them. Here contemporary philosophers face a certain difficulty, because they have all been trained to think of modality in terms of *possible worlds*. Possible worlds are often likened to *times*, so that just as we have questions of *diachronic* identity—identity over time—so we have questions of *transworld* identity, or identity across possible worlds. If one thinks of time as a dimension, with objects being located at different “places” in it, it may well be tempting to think of possibility as yet another dimension, involving another set of “places” at which objects may be “located.” Then, just as we want to formulate adequate criteria of identity *over time*, we shall feel an urge to formulate analogous criteria of transworld identity, telling us when an object located in one possible world is or is not identical with one located in another.

Let me say straightaway that I think that this approach to modality is simply *crazy*. I am not simply objecting to those, like David Lewis (1986), who think that possible worlds are spatiotemporally unified domains of concrete objects—effectively, *parallel universes* akin to the one that we inhabit. This is surely the stuff of science fiction—or, at best, of some highly speculative interpretations of quantum physics, the so-called many-worlds interpretations. I am also objecting to those possible-worlds theorists who qualify as what Lewis calls “ersatzists.” These are philosophers who regard possible-worlds as abstract entities of special kinds, such as maximal consistent sets of propositions. My most basic objection to all of these adherents of possible-worlds

accounts of modal truth is that it is a fundamental error to appeal to special *entities* of any kind to explain the nature and ground of modal truths, whether these entities be concrete or abstract in character. Why? Because there will be modal truths about entities of *any* kind, including “possible worlds,” if such entities exist. If possible-worlds exist, then the modal truths about *them* cannot be explained by the possible-worlds approach to modality, but will have to be explained in some other way. But then that other way will be able to explain all other modal truths too, rendering the whole apparatus of possible-worlds redundant. I shall explain in a moment what I take this “other way” to be, although I have already intimated what it is: it is the way of *essence*. But before moving on, however, let me just emphasize that I think that the whole notion of “transworld” identity is completely misguided, as is the attempted analogy between time and possibility.

3.2 Essence and Modality

When possible-worlds theorists talk about essence, as they often do, they explain such talk, as one might expect, in terms of their favored approach to modality. They say, for instance, that an *essential* property of an object is one that that object possesses “in every possible world in which it exists” and thus possesses *necessarily*—on the model, thus, of a *permanent* property of an object, this being one that it possesses at every time at which it exists. However, as Kit Fine (1994) has convincingly argued, this way of thinking of the relationship between essence and necessity is exactly back-to-front. Rather than attempt to explain essence in terms of necessity, we need to explain necessity in terms of essence. Apart from anything else, trying to explain essence in terms of necessity leads to all sorts of absurd or implausible essentialist theses. For instance, because it is true, on the possible-worlds account of necessity, that in every possible world in which Socrates exists, he is either a cabbage or such that 2 plus 2 equals 4, it will follow, on the standard account of essence, that it is an *essential* property of Socrates that he is either a cabbage or such that 2 plus 2 equals 4—and yet, very plausibly, that property has *nothing whatever* to do with Socrates’s essence or nature, that is, with what he *is*.

3.3 What are “essences”?

So what, then, do I mean by something’s *essence*? Locke put it well when he said that the word “essence” denotes “the very being of any thing, whereby it is, what it is” (1975 [1690], III, III, 15). Indeed, the word “essence” is just the English version of the Latin word that is standardly used to translate a phrase in Aristotle’s *Metaphysics* whose more literal translation is “the what it is to be” or “the what it would be to be.” So, as I have said earlier, a thing’s

essence is *what that thing is*—which is revealed by what metaphysicians in the Aristotelian tradition would call a *real definition* of the thing in question. When we grasp a thing's real definition, we grasp *what it is*, and this is what it is to grasp its *essence*. (It is important to remark here that *real* definitions are quite different from mere *verbal* definitions, which just provide synonyms for words or phrases.) In effect, grasping something's essence is simply *understanding what we are thinking about* in thinking of that thing. And this is surely something that we must, at least sometimes, be able to do, if we can think at all.

However, it is important to make a distinction at this point between a thing's *general* essence and its *individual* essence (the latter being what I called, in Part 1, its "identity," in one sense of that term). We grasp a thing's general essence when we know what *kind* of thing it is and we grasp its individual essence when, in addition, we know *which* thing of that kind it is (see Lowe 2008b). Having explained, now, what I mean by "essence," I want to go on to state and defend three fundamental principles or *laws of essence*, which are the following: (a) *essences are not entities*—we should not *reify* essences, (b) *essence precedes existence*—both ontologically and epistemologically, and (3) *essences are the ground of all modal truth*.

3.4 Essences Are Not Entities

Entities are things that *do or could exist*—and they are of many different kinds, belonging to many different ontological categories: material objects, persons, properties, relations, sets, numbers, propositions, . . . the list goes on indefinitely. Entities belonging to different ontological categories have different *existence and identity conditions*, the latter being expressed by the *criteria of identity* that govern the entities in question. An entity's existence and identity conditions, and hence its ontological category, are *part of its essence*—part of what it is to be the kind of thing that it is, and which thing of that kind it is. However, we should not think of an entity's essence as itself being a *further entity*, somehow specially related to the entity whose essence it is: this is to indulge in a mistaken *reification* of essence. (Note, however, that the *individuator* of any given entity, *x*—what *makes x* which thing it is—is an entity: and it is part of *x*'s essence that it is has this *individuator*: (see Lowe 2003, 2007b). So at least some of those who reify individual essences may well be confusing them with individuator.)

Since *all* entities have essences—there is always a fact of the matter as to *what* an entity is—if essences were entities, they would have to have their own essences, and so on *ad infinitum*, generating an infinite regress (see Lowe 2008b). Then, it seems, we could never know something's essence without knowing infinitely many other essences, making knowledge of essence impossible for finite minds like ours. In any case, treating essences as entities

makes knowledge of them problematic even if we ignore the regress problem, because it seems that they would have to be *esoteric* entities of some sort, not readily recognizable ones.

Locke himself fell into this error when he claimed that the “real essences” of material substances are their “unknown internal constitutions”—what we would now call their atomic or molecular structures. This error has been perpetuated in modern times in the form of the Kripke–Putnam doctrine that the essences of such substances are precisely such microstructures, which are discoverable only by advanced scientific techniques (see Putnam 1975 and Kripke 1980). It is true enough that sometimes—very often, in fact—knowing something’s essence involves knowing that it stands in a relation of essential dependence to some other thing or things. For instance, it is part of the essence of a *set* that it has the *members* that it does, so that we can only know *which* set it is if we know *which* things are its members. But that is not to say that its members *are* its essence, or even part of its essence: just that it is part of its essence *that* it has these members. (Note, however, that a set’s members are what *individuate* it—collectively, they are its individuator.) Philosophers in the *possible-worlds* tradition tend to reify essences in this way, regarding a thing’s essence as *the set of properties* that it possesses in every possible world in which it exists. But although it may be part of a thing’s essence *that* it has a certain property, *that property* cannot literally be a part of its essence—for the property is an *entity* and hence anything of which it is a part must likewise be an entity, which I have just denied that an essence can be.

3.5 Essence Precedes Existence

When I say that “essence precedes existence,” I mean this in both (a) an ontological sense and (b) an epistemological sense (see Lowe 2008b). In the first sense, what I mean is that a thing can exist only if its essence, together with essences of other existing things, *permits* its existence. Suppose we ask, for instance, whether a lump of bronze exists in a certain location, *L*. If a rock already exists in *L*, then the answer to our question is “No.” Why is that? Because, given *what a rock is* and *what a lump of bronze is*—given their *essences*—the presence of a lump of bronze in *L* is excluded by the fact that a rock is already located there, since rocks and lumps of bronze are *pieces of matter*, and it is part of the essence of any piece of matter that it excludes any other piece of matter from the place that it occupies. In the second sense of our expression, what I mean is that we can *know* something’s essence—know *what it is or would be*—before knowing *whether or not it exists*. This implies that knowledge of essence is, at least in a *relative* sense, a priori knowledge. I say “at least in a relative sense” because I want to acknowledge that very often we can only grasp new essences after having made empirical discoveries concerning the

existence and properties of other entities. A good example of this is provided by the transuranic elements, with atomic numbers greater than that of uranium. After atomic physicists had discovered and explored the nuclear structure of naturally occurring elements, they were able to predict the possible existence of these heavier elements—they knew in advance *what they would be*, but only because they already possessed considerable empirical knowledge of nuclear physics (see Lowe 2008b).

Incidentally, there is no conflict between what I am saying here and what I said earlier about Locke's views concerning the "real essences" of material substances and the modern version of this in the Kripke–Putnam theory. Kripke and Putnam claimed, for example, that water is *essentially* composed of H_2O molecules, but that this discovery was made empirically—so that, by their account, we did not really know *what water is* until relatively recently in human history. By "water" here they are referring to a certain *macroscopic kind of stuff*, which fills our oceans and falls as rain. But I can see no good argument for the claim that this kind of stuff is *essentially* composed of H_2O molecules, although, of course, I by no means deny that it is *actually* composed of such molecules (see Lowe 2008b). On the other hand, when I was speaking a moment ago about the "transuranic elements," I was referring to certain species of *atoms*, with certain combinations of protons and neutrons in their nuclei, not to macroscopic kinds of stuff. Putnam himself (1990), wisely enough, eventually abandoned his essentialist claim about water and H_2O , but many other philosophers continue to accept it uncritically. The claim rests on two highly dubious assumptions. The first is that it is part of the *general* essence of any macroscopic kind of stuff that it possesses its actual molecular structure, whatever that should turn out to be. The second is that from this first assumption, together with the empirical discovery that the *actual* molecular structure of water is H_2O , we can validly infer that it is part of the *individual* essence of water that it possesses the H_2O molecular structure. Inferences of this kind seem unproblematic within the possible-worlds framework, but are just as questionable as that framework is (see Lowe 2007a).

One reason why I am so insistent that, in the epistemic sense, "essence precedes existence" is that I believe that it makes no sense to suppose that we can really *think* about something if we are completely ignorant as to *what it is*. And yet this is the implication of the Kripke–Putnam version of a posteriori essentialism. Kripke and Putnam held, for instance, that although we have been thinking about cats for thousands of years, it could in principle turn out that, contrary to what we have always supposed, they are not living animals but are really robots sent from Mars. But what we *thought* we were thinking about when we supposed ourselves to be thinking about a *cat* was something with certain fairly specific *existence and identity conditions*—conditions that are *incompatible* with those of robots sent from Mars. I don't see how it can be the

case that we were *actually* thinking about something with the existence and identity conditions of such a robot. If we have been deluded for all these years, then what we *should* say is that we weren't really thinking about anything that *actually* exists at all—not that we were really thinking about something *quite different* from what we thought we were thinking about (see Lowe 2007b).

3.6 Essence is the Ground of all Modal Truth

This third and last law of essence should not need much explanation, given what I have already said earlier. The idea is that modal propositions are not made true by anything that does or could exist—actually existing entities or even pure “possibilia,” conceived as nonactual existents—but rather by the *essences* of things that do or could exist. These essences are, in accordance with the ontological version of the principle that essence precedes existence, ontologically *prior* to the entities whose essences they are and determine what is or is not possible regarding those entities. Moreover, in accordance with our first law of essence, these essences are not themselves *entities*, either actual or possible. Thus, for instance, it is owing to the *essences* of bronze statues and lumps of bronze that, unlike such a lump and a *rock*, such a lump and a bronze statue *can* exist in the same place at the same time. The upshot is that, in the logic of modality, the expression “it is part of the essence of *x* that . . .” should be taken as primitive, rather than the expression “it is necessary for *x* that . . .” or (worse still) “in every possible world in which *x* exists it is the case that . . .” As yet, a wholly satisfactory logic of essence and modality remains to be worked out. What I am convinced about, however, is that the possible-worlds approach to modality and essence must be replaced by an essence-based approach to modality.

3.7 Essentialism Is Not Conceptualism

A final point to emphasize is this. The kind of essentialism that I have been defending is not to be confused with any species of *conceptualism* (see Lowe 2008b). Essential truths are not grounded in *our concepts*. They concern, rather, the *mind-independent natures* of things. After all, concepts are just *entities* of a certain kind—*mental* entities, if one takes them, as I do, to be *ways of thinking of things* (see Lowe 2006, pp. 85–6). What is *really* possible or necessary cannot be grounded in *how we think about things*. Rather, our aim should be to accommodate our concepts to the real natures of things, for only then can we think adequately about reality. This, in large measure, is (or *should be*) the aim of philosophy and especially of metaphysics, as the greatest philosophers of antiquity—Socrates, Plato, and Aristotle—all took it to be.

Bibliography

- Burke, M. B., 1994. "Dion and Theon: An Essentialist Solution to an Ancient Problem." *Journal of Philosophy*, 91, pp. 129–39.
- Chisholm, R. M., 1976. *Person and Object*. London: George Allen and Unwin.
- Fine, K., 1994. "Essence and Modality." In J. E. Tomberlin, ed., *Philosophical Perspectives*, 8: *Logic and Language*, Atascadero, CA: Ridgeview Press, pp. 1–16.
- Geach, P. T., 1980. *Reference and Generality*. 3rd edn. Ithaca, NY: Cornell University Press.
- Kripke, S. A., 1980. *Naming and Necessity*. Oxford: Blackwell.
- Lewis, D. K., 1986. *On the Plurality of Worlds*. Oxford: Blackwell.
- Locke, J., 1975/1690. *An Essay Concerning Human Understanding*. Edited by P. H. Nidditch. Oxford: Clarendon.
- Lowe, E. J., 2009. "Serious Endurantism and the Strong Unity of Human Persons." In L. Honnefelder, E. Runggaldier, and B. Schlick, eds, *Unity and Time in Metaphysics*. Berlin: Walter de Gruyter, pp. 67–82.
- , 2008b. "Two Notions of Being: Entity and Essence." In R. Le Poidevin, ed., *Being: Developments in Contemporary Metaphysics*. Cambridge: CUP, pp. 23–48.
- , 2008a. "How are Identity Conditions Grounded?" In C. Kanzian, ed., *Persistence*. Frankfurt: Ontos Verlag, pp. 73–89.
- , 2007b. "Sortals and the Individuation of Objects." *Mind and Language*, 22, pp. 514–33.
- , 2007a. "A Problem for A Posteriori Essentialism Concerning Natural Kinds." *Analysis*, 67, pp. 286–92.
- , 2006. *The Four-Category Ontology: A Metaphysical Foundation for Natural Science*. Oxford: Clarendon.
- , 2003. "Individuation." In M. J. Loux and D. W. Zimmerman, eds, *The Oxford Handbook of Metaphysics*. Oxford: OUP, pp. 75–95.
- , 2002. *A Survey of Metaphysics*. Oxford: OUP.
- , 1998. *The Possibility of Metaphysics: Substance, Identity, and Time*. Oxford: Clarendon.
- , 1989b. "What is a Criterion of Identity?" *Philosophical Quarterly*, 39, pp. 1–21.
- , 1989a. *Kinds of Being*. Oxford: Blackwell.
- Olson, E. T., 2001. "Material Coincidence and the Indiscernibility Problem." *Philosophical Quarterly*, 51, pp. 337–55.
- Putnam, H., 1990. "Is Water Necessarily H₂O?" In *Realism with a Human Face*. Boston, MA: Harvard University Press, pp. 54–79.
- , 1975. "The Meaning of 'Meaning'." In *Mind, Language and Reality*. Cambridge: CUP, pp. 215–71.
- van Inwagen, P., 1981. "The Doctrine of Arbitrary Undetached Parts." *Pacific Philosophical Quarterly*, 62, pp. 123–37.
- Wiggins, D., 2001. *Sameness and Substance Renewed*. Cambridge: CUP.
- , 1980. *Sameness and Substance*. Oxford: Blackwell.

23 Philosophy of Mind: Consciousness, Intentionality and Ignorance

Daniel Stoljar

1 Introduction¹

In a striking passage at the end of his classic paper “Epiphenomenal Qualia,” Frank Jackson observed that “it is not sufficiently appreciated that physicalism is an extremely optimistic view of our epistemic powers. If it is true, we have, in very broad outline admittedly, a grasp our place in the scheme of things. Certain matters of sheer complexity defeat us . . . but in principle we have it all” (1982, 135). Jackson went on to suggest that, from a perspective that emphasizes that we are organic beings with an evolutionary history and limited psychological capacities this fact about physicalism renders it quite implausible.

It seems to me that Jackson’s instincts on target here, at least as regards the sort of materialism² he mostly had in mind, that is, the sort promoted by such philosophers as J. J. C. Smart (1959) and David Lewis (1983; 1994). However, what Jackson does not say is that the same thing is true of dualism. For dualism (at least in the form opposed by such philosophers as Smart and Lewis) agrees with materialism about all aspects of reality with one exception, viz., the particular states of mind having to do with consciousness. Moreover, while these states are an exception to materialism, they are not an exception to an extremely optimistic view of our powers, since the dualist typically supposes that the subjects who are in these states know what they are in what Lewis later called an “uncommonly demanding sense” (see Lewis 1995). Hence, just as it is not sufficiently appreciated that materialism is an extremely optimistic view, it is likewise not sufficiently appreciated that dualism is too. Indeed both the traditional positions in philosophy of mind—materialism and dualism—presuppose that “in principle we have it all.” If, following Jackson, we regard that presupposition as false (or is at least as incredibly unlikely) both positions should be rejected.

What happens to the philosophy of mind if we reject both these standard positions? I think that doing so puts us in a position to formulate new solutions to some of the central question of philosophy of mind, including the problem of consciousness and the problem of intentionality. The solutions do not involve what some philosophers say they want, a theory or account of consciousness or intentionality. But I think it is on reflection a mistake to expect solutions of this sort in any case, at least at the present stage of knowledge; the precise account of what consciousness or intentionality is and how they fit into the world are problems that will be solved if at all, by total science, by the science whatever it is that will be reached in the (perhaps hypothetical) end of inquiry. In the meantime, however, what philosophers certainly can do is assess whether certain lines of reasoning concerning consciousness or intentionality that lead to philosophically objectionable (or at any rate striking) conclusions are persuasive or not. When I say that rejecting the standard positions provides us with a solution to these problems, it is solutions of this latter sort I have in mind.

This chapter traces out how rejecting the standard positions leads to solutions of this kind. After providing some context for the issues, I start with the problem of consciousness, and briefly review my own favored epistemic solution to that problem—a solution hinted at but not developed in the passage from Jackson, and which I have set out elsewhere (see Stoljar 2006a; 2006b). Then, in the bulk of the chapter, I will consider how to extend this epistemic solution to the problem of intentionality. As I will explain, this is a nontrivial matter because the problem of intentionality is distinct from the problem of consciousness in not involving counterparts of the arguments distinctive of consciousness, such as the knowledge argument and the conceivability argument, and it is most obviously arguments of this style are subject to an epistemic response. So to see how to connect issues of ignorance to intentionality we will need to formulate the problem of intentionality in a more explicit way than it is usually done. In the brief concluding section of the chapter, I will make a remark about a third issue, the problem of self-knowledge, which is a problem that is different from the problem of consciousness and intentionality, but which has emerged as a key problem, perhaps the key problem, in philosophy of mind in recent years. I will suggest that while rejecting the standard positions does not solve this problem, the materials we assemble when thinking through the perspective suggested by Jackson are nevertheless important when we turn to this problem.

2 Responding to Our Inner Cartesian

One of the key events in philosophy of mind in the last 100 years—arguably, *the* key event—was the appearance in 1949 of Oxford philosopher Gilbert

Ryle's book *The Concept of Mind*. Ryle defined himself—and many philosophers of mind following him have likewise defined themselves—in opposition to someone called the “Cartesian.” The connection between Ryle's Descartes and the real Descartes is tenuous, and in any case isn't really to the point. Rather Ryle's Descartes acts as a foil in philosophy of mind rather as the skeptic about the external world acts as a foil in epistemology. The point is not to *refute a real person*, that is, someone who may or may not be the famous figure of seventeenth-century philosophy and science. The point is rather to *critically engage with an intellectual tendency*, which according to Ryle greatly influences our *own* contemporary philosophical interpretations of ordinary and scientific psychology, and which stands in the way of the attempt to provide a plausible general picture of human cognitive capacities in their relation to the rest of the world.

What exactly is this tendency? Ryle's Descartes holds two key theses. First, he holds the *metaphysical* thesis that mental phenomena and physical phenomena are wholly distinct; that is, he is a traditional dualist. Second, he holds the *epistemological* thesis that mental phenomena are wholly transparent to themselves; that is, each of us is equipped with a faculty of introspection that, if used properly, will provide us in principle with complete and infallible knowledge of the contents of our minds. Throughout *Concept*, Ryle develops a barrage of techniques and suggestions designed to undercut both theses. Nowadays many of these techniques and suggestions are unpopular, but the project he initiated—the project of accommodating our inner Cartesian, as we might put it—is still with us. Indeed, it is no exaggeration to say that the dominant focus of philosophy of mind since Ryle has been on the *first* of the two theses just distinguished.

The obvious way to resist being a traditional dualist is to become a traditional materialist, and indeed, this is the lesson that many philosophers took from Ryle. But Ryle himself thought that the traditional materialist holds a position that is almost as bad as the traditional dualist, since both hold a presupposition that is false. According to Ryle (at least as I read him³) the false presupposition at issue is that ordinary psychological declarative sentences are fact-stating just as ordinary physical declarative sentences are; that is, the sentence “Otto is anxious” states a fact just as “Otto is 6 foot 2 in his stockinged feet” states a fact. The dualist thinks that “Otto is anxious” states a fact, namely one about a realm only contingently connected to the physical. The materialist thinks that it states a fact about ordinary bodies. Ryle suggests that both sides are mistaken because “Otto is anxious”—and psychological declaratives generally—states no fact at all. In that sense both standard materialism and standard dualism are to be rejected.

However, it is difficult to make sense of the idea that “Otto is anxious” states no fact. For what does “state no fact” mean? A very common suggestion

is that a declarative sentence states no fact just in case an assertion of that sentence does not conventionally express a belief, and instead expresses a nonbelief, say a desire or intention or command. So, for example, the sentence “You will do your violin practice” might conventionally express a command, rather than a belief about the future. However, if this is what it is to be fact-stating, it is hard to believe that ordinary psychological sentences are not fact-stating; after all, does “Otto is anxious” not conventionally express the belief that Otto is anxious? What would constitute evidence that it does not? And anyway, to say that a sentence is conventionally used to express beliefs (or is not conventionally used to express beliefs) is *itself* a statement of fact about psychology, so, if it is true, Ryle’s position threatens to be self-undermining.

So in my view Ryle’s suggestion about things not being fact-stating does not represent a productive way to think about these issues. On the other hand, I do think he is right that both standard materialism and standard dualism presuppose something false. For they both presuppose an overly optimistic view of our powers. They both presuppose that “in principle we have it all.” But how does rejecting this presupposition allow us to respond to Ryle’s Cartesian? In the next section I will consider how to respond to the part of the Cartesian picture that most people think is the hardest to deal with: the problem of consciousness.

3 The Problem of Consciousness

One might have thought that it would be an easy matter to reject the first thesis held by Ryle’s Descartes, that is, metaphysical dualism. Surely a small dose of scientifically informed common sense—according to which humans are evolved creatures as much a part of the natural order as zebra fish or xenopus toads—is sufficient to dispel the idea that each of us is a complex of a body and soul, a picture apparently belonging more to the history of religion than to contemporary philosophical thought. As it turns out, however, things are not so simple. For as a number of philosophers of mind from the 1970s on argued (Nagel 1974; Kripke 1980; Jackson 1982; Robinson 1982; and Chalmers 1996) dualist modes of thought not only can be divorced from any religious element, but also can be founded on extremely compelling and simple intuitions about consciousness and then developed with considerable clarity using machinery borrowed from modern modal logic, semantics, and epistemology. To accommodate Cartesianism in this sophisticated modernized form requires us to ask some searching questions not only about our conceptions of nature and the mind but also about our conception of philosophical method.

One very plausible form of reasoning here has come to be called “the knowledge argument.” This argument may be set out in various ways, but a

simple version has it as proceeding from two main premises. The first premise concerns what it is possible for a person to know; in particular, it is possible for a person to know all the physical facts as well as every fact that follows a priori from the physical facts, and yet not know what it is like to have an experience of certain type. Jackson's (1982) Mary is the best known, but not the only, illustration of this possibility.⁴ The second premise of the knowledge argument is that if this is possible then materialism is false. The conclusion is that materialism is false, or anyway it is false if there are facts about what it is like to have certain experiences and if people know these facts.

How to respond to this argument? There are a number of existing proposals in the literature. One response, the ability response, is based on a distinction that Ryle himself developed and defended in *Concept*, namely the distinction between propositional knowledge, that is, knowledge attributed by sentences of the form "S knows *that* such and such," and know how, that is, knowledge attributed by sentences of the form "S knows *how to* such and such." Armed with this distinction proponents of the ability response say that while Mary learns something when she comes out of her room, she does not learn any propositional knowledge but rather gains a sort of know-how.⁵

But this proposal is problematic in two ways. First, it is not clear that there is any distinction here of the kind the ability response requires (see Stanley and Williamson 2001; Stanley 2011). Sentences of the form "S knows how to such and such" seem to be semantically similar to sentences such as "S knows where such and such is" or "S know who such and such is." But these sentences are usually thought of as attributing a sort of propositional knowledge; for example, when you know where something is, you know, for some suitably described place *t*, that the thing in question is at *t*. Likewise, if you know how to do something, it seems natural to say that you know, for some suitably described way of doing something *w*, that the thing in question is done in way *w*. If any analysis along these lines is correct, the ability hypothesis looks to be in serious trouble.⁶

The second problem is that, even if the know-how/know that distinction is granted, the ability response relies on the idea that the experience Mary has when she comes out of her room and sees color for the first time is a novel experience. The general idea is that since she has not had the experience she does not have the relevant abilities or know-how; similarly, when she does have the experience she will gain the abilities or know-how. However, it is possible to develop the knowledge argument on the basis of examples that do not involve novel experiences (see Stoljar 2005; 2006a). For such examples it is implausible that Mary gains a new ability—for she already has the ability in question (having already had the experience). If so, the ability hypothesis is not a good reply to the knowledge argument.

A different sort of response appeals to a distinction between two kinds of materialism—a priori and a posteriori materialism, as I will call them here.⁷ Suppose that we have a sentence “S” that somehow or other captures every physical fact of the world; and suppose we have a second sentence “S*” that somehow or other captured every psychological fact. Materialism of any sort is committed to the view that the material conditional formed from these sentences—that is, “If S then S*” —is necessary. The a priori materialist says that the conditional is necessary and a priori; the a posteriori materialist says that the conditional is necessary and a posteriori. Armed with this distinction, the a posteriori materialist response says that the knowledge argument shows at most that a priori materialism is false. But this leaves it open that a posteriori materialism is true.

But this response faces problems too. First, a number of philosophers insist that a posteriori materialism is not a possible position, however attractive it looks in the abstract. For these philosophers, there are theses in philosophy of language and epistemology that entail *if* materialism is true then a priori materialism is true (see, e.g. Lewis 1994; Chalmers 1996; Jackson 1998). These theses are contested of course, and we will not go in to these issues here (see, e.g. Stalnaker 2003). But the fact that many philosophers think that a posteriori materialism is not simply false but impossible at least shows that appealing to the necessary a posteriori is no easy response to the knowledge argument.

The second problem is that the a posteriori materialism response to the knowledge argument even if correct is likely to win the battle but not the war. For the knowledge argument is just *one* argument against materialism about consciousness. Other arguments, such as the conceivability argument, are not going to be defeated by drawing a distinction between necessary a posteriori, since those arguments focus directly on the question of whether the connection between the mental and the physical is necessary—the question of whether the connection is in addition a posteriori or not is from this point a view a sideshow. (A quick argument for this conclusion is that Kripke himself advanced something very much like the conceivability argument in *Naming and Necessity*, but also provided (in the same work) the materials for formulating a posteriori materialism. Kripke himself evidently did not think that a posteriori materialism would be able to answer the conceivability argument, which is good evidence, but of course not conclusive evidence, that a posteriori materialism will not answer the conceivability argument.⁸)

4 The Epistemic Response to the Consciousness Problem

If these common responses to the knowledge argument fail, how should one respond? In my view, the best response draws on the idea suggested in the passage from Jackson above.

Suppose that there is a type of nonexperiential or physical fact of which we are ignorant but that is relevant to the nature of experience. (In other work, I have called this supposition “the ignorance hypothesis.”) Now take the claim that it is possible that someone knows all the physical facts and not all the facts. Such a claim involves the phrase “all the physical facts,” which is what philosophers of language call “a quantifier phrase.” How should we interpret this phrase? Well if we suppose that the ignorance hypothesis is true, and there are physical facts of which we are ignorant but which are relevant, then we have two choices: to include them in the scope of the quantifier or not. Suppose first that they are included. Then it is not clear that we are entitled to assert the possibility claim. Is plausible to say that someone can know all the facts (including one’s of which we are ignorant) and yet not know all the facts? How can we assert that if we are ignorant of some of the relevant facts? Suppose now that they are not included. Now the possibility claim looks plausible; the physical facts that we currently know do indeed seem to be such that someone could know all of them and not know some phenomenal facts. But this possibility claim does not threaten materialism because at most it shows that facts about experience come apart from *some* physical facts not from them all. Putting this together, if the ignorance hypothesis is true, the knowledge argument is unpersuasive.

This response, if correct, will tell us that the knowledge argument is unpersuasive; to that extent the response is not committed to dualism. But one might reasonably ask why the position does not commit us to materialism, and if so, what happened to the point I mentioned before, that both standard materialism and standard dualism are to be rejected, just as Ryle thought (though not for Ryle’s reasons). Isn’t the epistemic response to the argument a straightforward response on behalf of the materialism?

The answer to this question is “no,” or better, “it depends on what you mean by materialism.” One way of spelling out materialism is to construe it as advancing a particular positive view about the physical world—for example, that contemporary physics (or something very like it) gives a complete statement of the world and what it is like. That is the sort of materialism that Lewis, for example defends, the sort I have so far called “standard materialism.” The epistemic response I have suggested is not available to a materialist of that sort, precisely because a materialist of that sort supposes that in principle we have it all. But one might use the term “materialism” in a nonstandard way, that is, to mean any position that supposes that the knowledge argument (and similar arguments) are unsound and that in consequence they provide us with no reason to suppose that facts about consciousness are fundamental. Using the label in that way, it is possible to say that materialism and the epistemic view are compatible. But that does not mean the epistemic view does not involve a rejection of standard materialism.

Of course the epistemic approach to the knowledge argument faces a number of challenges. Some argue that we are not ignorant in the way that the response requires; others argue that even if we are, then this will not have the effect on the arguments that the approach assumes it will. I have responded to these problems in detail elsewhere, and will not go over them here (see Stoljar 2006a). Instead, my aim in what follows is to consider whether the sort of response just sketched to the problem of consciousness may be applied to other problems in philosophy of mind.⁹

5 The Problem of the Problem of Intentionality

Traditionally, philosophers have distinguished the problems presented for materialism by phenomenal consciousness from those presented by another aspect of mental states, what contemporary philosophers—roughly following the nineteenth-century philosopher, Brentano—called their intentionality. The intentionality of a mental state is its aboutness. When I think of Vienna or I believe that the computer is on the desk or I fear that the planet will get hotter, and so on, I instantiate mental states that are in a hard to define sense about Vienna, or the computer on the desk or planet Earth. The idea is that mental states and events have a property rather like signs, sentences, and gestures; that is, they are about or represent things other than themselves.

Can we extend the epistemic response just suggested from the problem of consciousness to the problem of intentionality? Before we answer this question we need to be clear about what the problem of intentionality is. For, while it is certainly traditional to talk about the problem of intentionality, it is harder than you might think to formulate the problem in any precise way.

In the light of our discussion of the problem of consciousness, a natural first thought is that there is a problem of intentionality in just the same way as there is a problem of consciousness. So for example, just as we formulated a knowledge argument for consciousness, we might formulate a counterpart “knowledge argument” for intentionality. The first premise of this argument would say that it is possible for a person to know all the physical facts as well as every fact that follows a priori from the physical facts, and yet not know what people believe, (e.g.) not know that Hillary believes that Obama is president (assuming that Hillary does believe this). The second premise is that if this is possible then materialism is false. The conclusion is that materialism is false. If there were such a persuasive argument, we could formulate the problem of intentionality, on analogy with the problem of consciousness, as the problem of saying what if anything is wrong with this argument.

But the problem is that there is no intuitive foundation to this counterpart knowledge argument. The Mary case provides us with an initially plausible

case of someone who knows all the physical facts and yet does not know what it is like to see color. (That the case is not plausible on reflection does not mean it is not initially plausible.) But there seems to be no similarly plausible case of someone who knows all the physical facts and yet does not know that Hillary believes that Obama is President. Mary herself, for example, seems perfectly capable of knowing facts of this sort. If someone told her while in the room what Hillary believes, she might well come to know that she believes that Obama is President, and there seems no aspect of Hillary's belief of which she is ignorant.

It might be thought that while Mary may be able to come to know what Hillary believes in this way, there still seems to be gap between a complete description of the world in basic physical terms, and the existence of Hillary's belief. For example, if Mary were to read a description of Hillary's brain in basic physical terms, would she then be able to work out—from this description alone—what Hillary believes? If the answer to this is “no”—as it seems to be—then why is there not a knowledge argument for intentionality after all?¹⁰

However, this line of thought assumes that Mary's knowledge is limited to a certain sort of physical information about Hillary's brain, for example to cellular information. It is true that from this alone one cannot work out what Hillary believes. But Mary's knowledge is not limited in this way; indeed it is an important strength of the knowledge argument in its original form that Mary is permitted to know anything that she could come to know while in the room. So for example, suppose Mary has access not just to cellular information about Hillary but also to computational, functional, behavioral, and environmental information. If so, it is hard to believe that she could not come to know that Hillary believes that Obama is President, even if, as the original argument alleges, she could not come to know what it's like to see color.

Alternatively, it might be thought that there is an important class of intentional states that Mary could *not* come to know, viz., those that are also states of consciousness associated with experiencing color. On many views, after all, such states are *themselves* intentional states and so (one might think) it is quite an easy matter to produce a plausible case in which someone knows all the physical facts and yet does not know an intentional fact—perhaps the Mary case is precisely such a case.

However, while this might be true, it does not affect the basic issue. The problem of intentionality is supposed to concern intentional states as such, not merely intentional states that are also conscious states. Indeed, it is for this reason that it is natural to attempt to formulate the problem in terms of a standing belief such as the example of Hillary's belief. And if we focus on intentional states as such (i.e. those that are not also conscious states) then it seems there is no knowledge argument about intentionality.

If the problem of intentionality is not to be explained on a direct analogy with the problem of consciousness, perhaps it comes about from the idea that intentional properties are not fundamental? Here is Jerry Fodor forcefully giving voice to this idea:

I suppose that sooner or later the physicists will complete the catalogue they've been compiling of the ultimate and irreducible properties of things. When they do, the likes of spin, charm, and charge will perhaps appear upon their list. But aboutness surely won't; intentionality simply doesn't go that deep. It's hard to see, in face of this consideration, how one can be a Realist about intentionality without also being, to some extent or other, a Reductionist. If the semantic and the intentional are real properties, it must be in virtue of their identity with (or maybe of their supervenience on?) properties that are themselves neither intentional nor semantic. If aboutness is real, it must be really something else. (1987, 97)

The most obvious thing Fodor is saying in this passage is that intentionality is not a fundamental feature of the world. However, while this is plausible, it is again not sufficient to generate a problem about intentionality. For the same thing can be said about almost everything. For example, take the Mariana Trench. When the physicists complete the catalogue they've been compiling of the ultimate and irreducible properties of things, being the Mariana Trench surely won't be on that list either. The Mariana Trench goes deep but not that deep. But nobody thinks that there is a philosophical problem about the Mariana Trench. Put differently: if there is a problem of intentionality at all, it had presumably have something to do with intentionality, but the fact that intentional properties are not fundamental is a fact shared by many things.¹¹

At this point a third suggestion about what the problem of intentionality is naturally suggests itself. Intentionality is usually associated with a number of interesting logical features, and the problem of intentionality might be thought of as the problem of coming to grips with these features. Suppose for example, that Hillary believed, not that Obama is President, but that Odin is. In that case she would have a belief about Odin rather than a belief about Obama. But off-hand this is puzzling. To have a belief about Obama might be thought of involving a relation between Obama and the person who has the belief—Hillary in this case. But the example of Odin shows that this cannot be so, or at least cannot be so in general. Hillary cannot stand in any relation to Odin for the simple reason that Odin does not exist. How then can she believe that Odin is President?

However while it is certainly challenging to explain the sense in which Hillary can have a thought about Odin, it is doubtful that we can straightforwardly

appeal to this to raise the problem of intentionality, at any rate not if we are out to formulate a problem that has something to do with the contrast between materialism and dualism. For the Odin problem has nothing to do with the truth or not of materialism! Suppose for example that dualism of some very straightforward kind is true, and that Hillary was a complex of a soul and a body—it would *still* be puzzling how something that exists can apparently stand in a relation to something that does not. Similarly, suppose that it were a fundamental fact of nature that Hillary believes that Odin is President—even so, it would remain the case that the fact in question cannot be analyzed as relation between Hillary and Odin, for there is no such relation.

In summary, before we confront the problem of intentionality, we had better confront the problem of the problem of intentionality, that is the problem of saying what the problem of intentionality is. As we have seen, is not generated by a counterpart for intentionality of arguments like the knowledge argument—for there is no such argument. It is not generated by the fact that intentionality is not a fundamental feature of reality—for while it plausibly is not fundamental, this is not a feature of intentionality in particular. And it is not generated by the fact that intentionality exhibits interesting logical features—for while these features need explanation, doing so seems to have nothing to do with the distinction between dualism and materialism. What then *is* the problem?

6 The Descriptive/Foundational Distinction

In my view, a good way to proceed here¹² is to start with a distinction that is well known in the philosophy of language but is often less explicitly drawn in the philosophy of mind. This is a distinction, in Robert Stalnaker's (1997) terms, between questions of descriptive, and questions of foundational, semantics. The descriptive semantic project, as Stalnaker describes it, concerns what language we as a community speak, or if we confront the issue in an individualist framework, what language a particular individual speaks. Suppose, to fix ideas, we focus on a particular person—Karl, for example, the hero of David Lewis's (1974) paper "Radical Interpretation." The descriptive semantic project with respect to Karl is then the project of saying what language it is that he speaks.

This question about Karl is an empirical question. It might be that he speaks some particular dialect or idiolect of English, or it might be that he speaks some particular dialect or idiolect of Urdu. (That the answer to this descriptive semantic question may seem obvious does not mean it is not empirical.) Likewise, it might be that he speaks a language in which names are semantically equivalent to definite descriptions or that he speaks a language in which

names are not semantically equivalent to definite descriptions. (That the answer to this descriptive semantic question may seem unobvious does not mean it is not empirical.)

Now if the descriptive semantic project is empirical, to solve it we need to attend to various sources of evidence that we have about Karl. How he acts in particular circumstances is surely one good source of evidence, and as are the judgments he (and we) would make about when certain sentences are true and under what conditions. It might also be that other sorts of evidence—say about what sort of creature Karl is—are also relevant. Indeed, in principle anything at all can be evidence for the hypothesis that Karl speaks a particular language; all that is required is that the evidence, together with background assumptions, makes the hypothesis more probable than it would otherwise be.

In saying that the question of what language Karl speaks is an empirical question, I am not denying that there are *a priori*, or at least very general, constraints on what it is for Karl or anybody to speak a language, and that these general constraints will also factor into descriptive semantics. For example, one condition mentioned by Stalnaker is this: “if the semantics is correct, then speakers must know, at least for the most part, what according to the semantics they are saying” (1997, 176). As I understand it, a condition like this functions to narrow down the possibilities of what language it is that Karl is speaking. *A priori*, after all, the possibilities are endless and the evidence that we have about Karl will surely underdetermine which language he speaks. But the assumption that Karl must be assumed to know what he is saying, at least for the most part, serves to narrow down the possible languages which—we can reasonably suppose—Karl speaks.

Suppose now that after reviewing the evidence and the relevant general constraints we agree that Karl speaks a particular language—L17 as it might be. Then we face what Stalnaker calls foundational semantic questions; these are questions about “what the facts are that give expression their semantic values, and more generally, about what makes it the case that the language spoken by a particular individual or community is a language with a particular descriptive semantics” (1997, 167). Concerning Karl, then, questions at the foundational level concern what facts about him—for example what psychological, physical, or behavioral facts—explain that he speaks L17 rather than something else or nothing at all. Presumably for example there are facts about Karl that make it the case that L17 is a version of English rather than what we call Urdu (if it is). And presumably too there are facts about Karl that make it the case that L17 is a language that contains names not equivalent to any definite description (again: if it is). Questions at the foundational level ask what exactly these facts are.

7 From Language to Mind

Now, as the reference to semantics makes clear, the descriptive/foundational distinction is in the first instance a distinction in the philosophy of language. But the problem of intentionality if it is anything is a problem in the philosophy of mind. So, to connect this distinction to the problem of the problem of intentionality we would need to transpose it from the key of language to the key of mind. How is this transposition to be achieved?

One proposal might be to suggest that Karl has a language of thought, and then to apply the distinction directly to Karl's language of thought. Descriptive semantics for the language of thought from this point of view proceeds just like descriptive semantics for English or Urdu. However, formulating the problem of intentionality this way seems to me a mistake. The problem is not that Karl does not have a language of thought. Arguments such as those given by Fodor (e.g. Fodor 1987) seem to me quite compelling; they provide good empirical evidence (though of course not conclusive evidence) that Karl has a language of thought, or at any rate would if (as we are assuming) he is a human being like the rest of us. The problem is rather that the hypothesis that Karl has the language of thought is a quite specific psychological hypothesis about him; much more specific, for example, than the hypothesis that he has a mind or is the subject of intentional states at all. On the face of it, it is possible (even if not actual) that Karl could have intentional states without having a language of thought. But if that is so, we do not want to explain what the problem of intentionality is in terms of the language of thought.

A better proposal is to draw a distinction between (as I will say) descriptive psychology and foundational psychology (i.e. on direct analogy with descriptive and foundational semantics).¹³ Descriptive psychology is the project of saying what mental states Karl has; for example, what states or systems of states of knowledge, belief, desire, feeling, perception, imagination, memory he has. As in the case of descriptive semantics, the questions one raises concerning descriptive psychology are empirical questions, and so will need to be responsive to some sort of evidence. Once again, our evidence here will surely include how Karl acts in particular circumstances. But we also might include evidence about how similar Karl is to us, about what sort of creature he is and so on. For example, if Karl stubs his toe, and jumps around, it would be natural for us to attribute to Karl pain of the sort that we would feel if we were in the same sort of situation.

In saying that the question of what mind Karl has is an empirical question, I am as before not denying that there are a priori, or at least very general, constraints on, or theses about, what it is for Karl to have a mind, and that these general constraints will also factor into descriptive psychology. For example, it seems reasonable to suppose that if Karl has some belief states, then together

with other states that will cause him to act in certain ways. But the suggestion that belief states have causal powers seems to be a claim about what beliefs are in general, rather than a specific claim about what beliefs Karl has.

Suppose now that after reviewing the evidence and the relevant general constraints we agree that Karl has a particular mind—M17 as it might be. Then, in parallel with the language case, we face foundational psychological questions; these are questions about what facts about Karl make it true that he has M17 as opposed to some other mind, or opposed to no mind at all. If one is a dualist one might well say that Karl's having M17 is a fundamental fact, whereas if one is not a dualist one will say that there are other facts about him in virtue of which he has M17. The foundational project is to say what those facts are.

8 The Intentionality Problem and Descriptive Psychology

We have reviewed the descriptive/foundational distinction, and proposed a way to extend that distinction to the philosophy of mind. But how does this help with formulating the problem of intentionality, with "the problem of the problem of intentionality" as I called it earlier?

Well, it is very common, in the light of this distinction, to say that the problem of intentionality is a problem about foundational rather than descriptive semantics. Transposing this to the key of mind, the suggestion is that the problem of intentionality is a problem about foundational rather than descriptive psychology. I think there is something right about this suggestion, but it is also misleading. The reason it is misleading is that a lot of the issues that philosophers of mind discuss when they discuss the problem of intentionality turn out to be on the descriptive side of the divide. I will give four examples.

First, take the problem of thinking about nonexistent things, such as Odin, that we considered a moment ago. Suppose that the mind that we attribute to Karl—M17—includes the belief that Odin likes ravens. If so, we cannot construe this belief as involving a relation between Karl and the subject of his beliefs, that is, Odin. For Odin does not exist; hence it is impossible for Karl to stand in a relation to him. This is certainly an aspect of the problem of intentionality, but as we noted before it has very little to do with the contrast between dualism and materialism. In the light of the descriptive/foundational distinction it seems fairly clear why: the problem is one of descriptive psychology. To see this, notice that a common way to solve this problem is to say that the mind that Karl has (i.e. M17) must somehow involve a relation between Karl and an abstract object, for example the property of being the king of the Gods, who plucked out one eye to gain infinite wisdom, who has an eight-legged horse, etc. This abstract object exists but is not instantiated, that is

because Odin does not exist. We might also want to say that if Karl has beliefs about Obama this too involves a relation to an abstract object, it is just that in this case the abstract object is instantiated, that is, in Obama. Making these assumptions about the mind that Karl has raises further issues—how a concrete object can stand in relation to an abstract object, for example. I will not go into that here. The point is that the hypothesis that Karl stands in a relation to various abstract objects seems to be something we arrive at through theorizing about what sort of mind he has, and so through descriptive psychology.

Second, take the dispute in the philosophy of perception over whether perceptual states are relations to concrete objects, as emphasized by disjunctivists, or whether they involve representational states of some sort.¹⁴ This is an aspect of a question about intentionality too, but again it is a question in descriptive psychology. For example, if we attribute M17 to Karl, this will certainly involve some facts about perception, and about how perception relates to belief and so forth. What is the nature of perception? The representationalist says that perceptual states are in some ways akin to belief states in that they involve a certain kind of representational state. Suppose for example that M17 is a mind that involves a certain kind of representational state of the sort mentioned by representationalists, while M92 involves no such state, and simply says that Karl bears a phenomenological relation to his surroundings. Both hypotheses are plausibly compatible with various sorts of data, for example, behavioral data and introspective data, but they are different from each other. The disjunctivist thinks that the problem with saying that Karl has M17 is that there is no M17 to have, that is, for there are no perceptual representational states at all. The representationalist, by contrast, thinks that is not so, and that M17 is a possible mind. If so, there is no problem with saying that Karl has M17.

A third example concerns principles of charity, as discussed famously, for example, by Donald Davidson (see Davidson 1974). According to him, when we attribute M17 to Karl, we should be driven by the *a priori* principle that most of Karl's beliefs are true, and presumably that most of Karl's perceptual states are veridical. Suppose the hypothesis that Karl has M17 entails that most of his beliefs are true, while the hypothesis that Karl has M45 entails that most of his beliefs are false. Davidson's principle of charity is that it is constitutive of the nature of belief that most of a person's beliefs are true. So while M17 might be equivalent to M45 in respect of behavioral evidence, it is rational for us to adopt the hypothesis that M17 is Karl's. Other philosophers disagree with Davidson here, arguing that we have no *a priori* reason to favor M17 over M45. I don't want to engage this dispute but to note only that it is a disagreement about descriptive psychology.

As a final example, take the dispute about naturalness. There is a problem famously posed by Kripke (following Wittgenstein) about whether Karl—to

adapt the issue to our own discussion—is adding or quadding, where to quad two numbers is to produce their sum up to some limit, and then to produce 5 thereafter (see Kripke 1982). In his discussion of these matters, Lewis says that that in this case it one should adopt that view that Karl is adding, rather than quadding, because (to put it roughly) this is the most natural rule that Karl could be following (see Lewis 1983). As I understand it, it is conceded by Lewis that our evidence either from introspection or from behavior could not discriminate these hypotheses; the fact that one hypothesis is natural is suggested as a further constraint that could. Lewis's suggestion is controversial, but for the moment I am not interested in assessing it. Instead, I am interested in noting that it seems to be a part of descriptive psychology. In particular, Lewis seems to be suggesting that in developing our theory of Karl we would need to be driven by the a priori constraint that Karl's mental states attitudes are likely to be natural ones, other things being equal.

In sum, it turns out that a lot of the questions that philosophers discuss when they discuss intentionality are questions in descriptive psychology. This point is important because it shows that the point about the problem of intentionality we mentioned at the beginning of this section—that it concerns foundational rather than descriptive questions—is at best half right. But it is also important for another reason, and this has to do with the question we raised earlier, viz., whether the epistemic response to the knowledge argument might be extended to the problem of intentionality. If by “the problem of intentionality” we mean what we might call the descriptive problem of intentionality—that is the (complex) problem of saying what mind Karl has—then it would seem that there is no easy extension of the epistemic response to the problem of intentionality. This is not to say that we have nothing to learn about what mind Karl or anyone has—on the contrary, the questions here (as we have seen) are empirical, and with respect to those questions, the best policy is surely “tolerance and the experimental spirit,” as Quine famously said. Nevertheless, it is not as if when we engage in the project of descriptive psychology we are concerned to assess arguments like the knowledge argument whose persuasiveness depends on all the facts being in; rather the issues have a different shape entirely.

9 The Intentionality Problem and Foundational Psychology

We have seen that if the problem of intentionality is interpreted as part of the project of descriptive psychology, then the epistemic response to the problem of consciousness is of only marginal relevance to it. But suppose the problem is interpreted instead as part of the project foundational psychology. At the foundational level, we face the question of in virtue of what (if anything) Karl

has M17. To focus on a specific mental state, suppose that, as part of having M17, Karl believes that Obama is President. With respect to this belief, the foundational question we need to focus on is this: in virtue of what does Karl have this belief?

In my view, it is at this point that the considerations we marshaled in the course of developing the epistemic approach to the problem of consciousness have a role to play we turn to thinking about intentionality. The reason is that it is possible to sketch an answer to the question just posed, and this answer is more plausible than it would otherwise be if considered in the light of the epistemic approach.

The answer I have in mind is a version of the well-known Lewis–Armstrong argument for the identity theory (see Lewis 1966; Armstrong 1968). Transposed to our discussion, the first premise of this argument is that when Karl believes that Obama is President he is in a state that plays a particular theoretical role—that is, it is a state that in Karl produces other states, and is produced in such and such circumstances, and produces such and such actions, etc. We might summarize this by saying that according to the first premise the belief that Obama is President in Karl is that state which satisfies role R. The second premise of the argument is that there is some physical state of Karl that satisfies role R. The conclusion drawn from these two premises is that the belief that Obama is President is that physical state. If that argument is sound, it would be fair to say that we would have answered the foundational question about Karl, namely by saying in virtue of what he believes that Obama is President, namely in virtue of being in that physical state.

However, while this argument would (if sound) answer the foundational question, it raises a number of complicated and difficult issues. The first concerns the first premise of the argument. Lewis's defense¹⁵ of this premise involves the suggestion that the premise is not simply true, but true by definition; that is, Lewis thinks that the state of believing that Obama is President may be defined as the state whatever it is that plays the relevant role, something that follows simply from an understanding of the terms. Moreover, according to Lewis, the definition in question (a) constitutes a reductive definition in the sense that the role itself may be spelled out using no psychological vocabulary at all, and (b) is tacitly known by us, somewhat in the way that we know the syntactic rules of our native language (see Lewis 1994).

But, in view of the controversy surrounding the possibility of reductive definitions in philosophy of any sort, this defense of the first premise makes the overall argument seem less plausible than it otherwise might be. Moreover, it is an assumption that is not required by the soundness of the argument: for a valid argument to be sound all that is required is that the premises are true, not that one of them is true by definition. In the light of this, a very natural

suggestion is that Lewis–Armstrong argument is much more plausible if assumption that its first premise is analytic is dropped.

However, before we agree to this suggestion, we need to confront the reasons Lewis has for supposing that the first premise is analytic. I think there are a number of considerations motivating Lewis at this point but perhaps the main one (and the one I will concentrate on) is that when he advances this argument, Lewis is concerned to defend, not simply the identity of mental states with physical states, but a certain sort of a priori materialism—that is, the position we contrasted with a posteriori materialism earlier. In particular, he is interested in the idea that, if the first premise of the argument is true by definition then the second premise of the argument will provide a physical statement that a priori entails the conclusion. And this is exactly as a priori materialism requires.

However, in the light of the epistemic approach to the problem of consciousness mentioned above, it should be clear that is not necessary to defend a priori materialism in this way. For, in the light of that approach, it is possible to separate out two distinct claims: the first claim is that the physical facts, whatever they are, a priori entail the psychological facts; the second claim is that it is possible to define the psychological facts in terms of the physical facts that we currently understand. The epistemic approach is not opposed to the first claim (though it is not committed to it either); that is, it is not inconsistent with that approach that the physical facts a priori entail the psychological facts. But it *is* opposed to the second claim, that is, because according to it we are ignorant of some of the physical premises required in the entailment. Hence if the picture associated with the epistemic response is coherent, then the first claim of the two just distinguished may be true even if the second is not.

How does this distinction make it more plausible to deny that the first premise of the Lewis–Armstrong argument is true by definition? Well, as we just saw, Lewis's reason for supposing that the first premise is true by definition is that *if this is so, then* a priori materialism will be true. And Lewis is undoubtedly correct in asserting this conditional claim. On the other hand, in the light the distinction just made, the reverse conditional is not true: a priori materialism might be true, *even if* the first premise of the argument is not true by definition—indeed, that is possibility made salient by the epistemic view. But this allows us to agree with Lewis that a priori materialism is true but disagree with him that the first premise of the argument is true by definition. And this in turn makes the argument much more plausible than it would otherwise be.

The assumption that its first premise is true by definition is one controversial feature of the Lewis–Armstrong argument. Another concerns its second premise, the suggestion that there is some physical state of Karl that plays the

relevant role. The usual way to motivate this premise is to say that materialism is true, and hence that there must be some state which plays the role (if the role is played at all). However, if the materialism at issue here is the sort we referred to earlier as “standard materialism” this premise seems implausible. For if standard materialism is true, then the premise says that there is some physical state *of a type currently known* of Karl that plays role R, and so, is his believing that Obama is President. But this seems to greatly overstate the current level of understanding that we have into matters of this sort. In some cases, it is plausible to think that the relevant sciences here—that is, cognitive psychology and neuroscience—have progressed to the point where they might identify some computational state of the brain with which particular mental states might be identified. But in many cases this is not plausible: “The current situation in cognitive science is light years from being satisfactory. Perhaps somebody will fix it eventually; but not, I should think, in the foreseeable future, and not with the tools that we currently have at hand” (Fodor 2000, 5). In short, if the second premise is understood in the light of standard materialism, one might well reject it on empirical grounds. In turn, however, to reject it on empirical grounds is to give up the idea that there is any physical state in virtue of which Karl believes that Obama is President; at this point, the dualist alternative seems the only option.

However, in the light of the epistemic approach to the problem of consciousness mentioned, it should be clear that it is not necessary to defend the second premise of the argument by appealing to standard materialism. For suppose instead we operate with nonstandard materialism, suppose the argument is set against the backdrop of nonstandard materialism, that is, against the view that tolerates the idea that we are missing certain types of facts that are relevant to the nature of mind. Then we can think of the Armstrong–Lewis argument as setting out a strategy for solving the problem of intentionality, not as an argument that we currently have the materials to complete. In summary, the perspective suggested by the passage from Jackson from which we began suggests that the premises of the Lewis–Armstrong argument can be defended in a different way from that suggested by Lewis. In turn, doing that provides us with a better answer than we might otherwise have to questions distinctive of foundational psychology.

10 The Problem of Self-knowledge

We noted at the outset that Ryle’s Cartesian holds two theses. The first is the metaphysical thesis that mental phenomena and physical phenomena are distinct. The second is the epistemological thesis that the mind is transparent to itself, that is, that we have an introspective faculty which if used correctly

can in principle illuminate all aspects of our mind. We have been concentrating on problems and arguments involved in the assessment of the first thesis, suggesting that the epistemic view is sufficient to answer arguments associated with consciousness, and helps out with the arguments associated with intentionality.

Turning to the second thesis, as in the case of metaphysical dualism, one might have thought that a small dose of scientifically informed common sense would be sufficient to reject it as well. Certainly it is a common feature of our intellectual culture that people are in many ways (as psychologist Timothy Wilson has put it recently) *strangers to themselves* (see Wilson 2002). Social psychologists (not to mention many modern novels) routinely tell us that we are often quite wrong about our own basic motives, desires, and character traits. Similarly cognitive psychologists and neural scientists portray the human mind as a congeries of different sub-systems operating independently of each other and on principles that are largely unknown to us (e.g. Fodor 1983). From this point of view, it is difficult to believe the picture of the mind as an arena in which, in principle, nothing is hidden.

As in the case of metaphysical dualism, however, things are not so simple. For what has emerged particularly in recent discussions (e.g. Alston 1971; Shoemaker 1994; Wright 2000; Moran 2001; Byrne 2005) is that while we certainly do not have privileged access to *all* of our mental states, it is nevertheless the case that at least for *some* mental states our first-person knowledge is quite *different* in character from (even if not always *better* than) the knowledge that one might have of the conscious states of others or indeed of other things quite generally. Moreover, this modernized form of privileged access has proved difficult to formulate precisely and leads to a number of puzzles and questions, just as Ryle thought, puzzles that have emerged in a somewhat piecemeal form in the literature over the last few decades of philosophical writing.

How might one explain the sense in which self-knowledge is different from other knowledge? It is obviously too late in the chapter to give this question adequate attention. But I think that the discussion we have been having about how to think about the problem of intentionality permits us to make a remark about how to think about the self-knowledge problem too. For the problem of self-knowledge is fruitfully thought of as a problem of descriptive psychology, rather like the problems about nonexistence, representationalism about perception, principles of charity, and naturalness that we considered earlier. We noted before that, in the case of descriptive semantics, it seemed reasonable as an *a priori* principle that if some descriptive semantic theory of the language that Karl speaks is correct, then Karl must know, at least for the most part, what according to the semantics he is saying. A parallel suggestion, though suitably modified, might be true in the case, not of language but of mind:

if some descriptive psychological theory of Karl is correct, he must be able to know, at least for the most part and to the extent that he is rational, what mental states he is in according to the theory. From this point of view, Karl, for example, cannot view what mental states he is in (or what his words mean) as a subject matter that he may or may not take an interest in—in contrast, say to Russian literature, which surely is a topic he may or may not take an interest in. Rather it is a subject matter about which if he is rational he can be assumed to have a certain sort of potential expertise. Spelling out what this expertise amounts too is a difficult matter, and will need to be left for another occasion. The point for us is that it is a project in descriptive psychology.

Notes

- 1 Thanks to the editors of this volume, Barry Dainton and Howard Robinson, for comments on a previous draft.
- 2 I will use the phrase ‘materialism’ in this chapter rather than ‘physicalism,’ which is used here by Jackson. Nothing turns on this, though the history of these words is of some interest. For discussion see ch.1 of Stoljar 2010.
- 3 In *Concept*, Ryle says that sentences such as “John Doe knows French” are “neither reports of observed or observable states of affairs, nor yet reports of unobserved or unobservable states of affairs.” (1949, p. 120). I take it that what Ryle meant by this is that expressions such as “John Doe knows French,” and presumably other psychological reports, are not in the fact-stating business.
- 4 Here is Lewis’s very vivid description of the case: “Mary, a brilliant scientist, has lived from birth in a cell where everything is black and white. (Even she herself is painted all over.) She views the world on black-and-white television. By television she reads books, she joins in discussion, she watches the results of experiments done under her direction. In this way she becomes the world’s leading expert on color and color vision and the brain states produced by exposure to colors. But she doesn’t know what it is like to see color. And she never will, unless she escapes from her cell” (1988, 263).
- 5 For the classic defense of the ability hypothesis see Lewis 1988, and the references therein.
- 6 I do not say that this objection is conclusive; it depends if there is a fallback position for the ability response. For some discussion of this see Stoljar (forthcoming).
- 7 In Chalmers 1996, these positions are referred to as “Type-A materialism” (= a priori materialism) and “Type-B materialism” (= a posteriori materialism).
- 8 For my own development of the issues raised in this paragraph, see Stoljar 2006a; 2006b.
- 9 In Stoljar 2006a, I was very wary of extending the epistemic response to consciousness to other problems. The present chapter represents a slight (but only slight) softening of that position.
- 10 Thanks to Barry Dainton for this objection.
- 11 For some literature in which points like this are emphasized, see Stich 1992 and Tye 1992.
- 12 A rather different way to proceed is to connect the intentionality issue more directly with consciousness, as suggested for example in Johnston 2007 and Pautz 2010. I will not try to engage directly with these interesting ideas in what follows.

- 13 In his 1997 paper, Stalnaker agrees (as I read him) that the descriptive/foundational distinction has a counterpart in the philosophy of mind, but does not go into detail about what that counterpart is, beyond noting (as I have done) that appealing directly to the language of thought is a mistake. See also Stalnaker 2004.
- 14 For the literature on disjunctivism, see the papers in Byrne and Logue 2009, and the paper by Paul Snowdon in this volume.
- 15 Armstrong seems to me more equivocal on the matter of issue of whether the first premise of the argument discussed in the text is true by definition, and so I will concentrate on Lewis here.

Bibliography

- Alston, W. P., 1971. "Varieties of Privileged Access." *American Philosophical Quarterly*, 8 (July), pp. 223–41.
- Armstrong, D. M., 1968. *A Materialist Theory of the Mind*. London: Routledge.
- Byrne, A., 2005. "Introspection." *Philosophical Topics*, 33, pp. 79–104.
- Byrne, A. and Logue, H., 2009. *Disjunctivism*. MIT Readers in Contemporary Philosophy. Cambridge, MA: MIT Press.
- Chalmers, D., 1996. *The Conscious Mind*. New York: OUP.
- Davidson, D., 1973. "Radical Interpretation." *Dialectica*, 27, pp. 313–28.
- Fodor, J., 2000. *The Mind Doesn't Work that Way*. Cambridge, MA: MIT Press.
- , 1987. *Psychosemantics*. Cambridge, MA: MIT Press.
- , 1983. *The Modularity of Mind*. Cambridge, MA: MIT Press.
- Jackson, F., 1998. *From Metaphysics to Ethics*. Oxford: OUP.
- , 1983. "Epiphenomenal Qualia." *Philosophical Quarterly*, 32, pp. 127–36.
- Johnston, M., 2007. "Objective Mind and the Objectivity of Our Minds." *Philosophy and Phenomenological Research*, 75 (2) (September 2007), pp. 233–68.
- Kripke, S., 1982. *Wittgenstein on Rules and Private Language*. Cambridge, MA: Harvard University Press.
- , 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.
- Lewis, D., 1995. "Should a Materialist Believe in Qualia?" *Australasian Journal of Philosophy*, 73, pp. 140–4. Reprinted in his 1999 *Papers in Metaphysics and Epistemology*. Cambridge: CUP, pp. 325–31. All references are to the reprinted version.
- , 1994. "Reduction of Mind." In S. Guttenplan, ed., *A Companion to the Philosophy of Mind* (Oxford: Blackwell, 1994), pp. 412–31. repr. in his *Papers in Metaphysics and Epistemology*. Cambridge: CUP, 1999, pp. 291–324. All references are to the reprinted version.
- , 1988. "What Experience Teaches." *Proceedings of the Russellian Society*, 13, pp. 29–57, repr. in his *Papers in Metaphysics and Epistemology*. Cambridge: CUP, 1999, pp. 325–31. All references are to the reprinted version. References are to the reprinted version.
- , 1983. "New Work for a Theory of Universals." *Australasian Journal of Philosophy*, 61, pp. 343–77, repr. in his *Papers in Metaphysics and Epistemology*. Cambridge: CUP, 1999, pp. 8–55. All references are to the reprinted version.
- , 1974. "Radical Interpretation." *Synthese*, 23, pp. 331–44.
- , 1972. "Psychophysical and Theoretical Identifications." *Australasian Journal of Philosophy*, 50, pp. 249–58. Reprinted in his 1999 *Papers in Metaphysics and*

- Epistemology*. Cambridge: CUP, pp. 248–61. All references are to the reprinted version.
- , 1966. “An Argument for the Identity Theory.” *The Journal of Philosophy*, 63, pp. 17–25.
- Moran, R., 2001. *Authority and Estrangement*. Princeton University Press.
- Nagel, T., 1974. “What It Is Like to Be a Bat?” *Philosophical Review*, 83, pp. 435–50.
- Pautz, A., 2010. “A Simple View of Consciousness.” In G. Bealer and R. Koons, eds., *The Waning of Materialism*. Oxford: OUP, pp. 25–67.
- Robinson, H., 1982. *Matter and Sense*. Cambridge: CUP.
- Ryle, G., 1949. *The Concept of Mind*. Routledge and Kegan Paul. Reprinted 1963; references are to the reprinted version.
- Shoemaker, S., 1994. “Self-knowledge and ‘Inner Sense’.” *Philosophy and Phenomenological Research*, 54, pp. 249–314.
- Smart, J. J. C., 1959. “Sensations and Brain Processes.” *Philosophical Review*, 68, pp. 141–56.
- Snowdon, P., 2013. “The Philosophy of Perception.” In B. Dainton and H. Robinson, eds, *The Bloomsbury Companion to Analytic Philosophy*, Chapter 27.
- Stanley, J., 2011. *Know How*. Oxford: OUP.
- Stanley, J. and Williamson T., 2001. “Knowing How.” *Journal of Philosophy*, 98, pp. 411–44.
- Stalnaker, R., 2003. “Conceptual Truth and Metaphysical Necessity.” In R. Stalnaker, 2003, *Ways a World Might Be*. Oxford: Clarendon, pp. 201–16.
- , 2004. “Lewis on Intentionality.” *Australasian Journal of Philosophy*, 82 (1), pp. 199–212.
- , 1997. “Reference and Necessity.” In Wright and Hale, eds, *Blackwell Companion to the Philosophy of Language*. Oxford: Basil Blackwell, pp. 534–54. Reprinted in R. Stalnaker, 2003. *Ways a World Might Be*. Oxford: Clarendon, pp. 165–87. References are to the reprinted version.
- Stich, S., 1992. “What Is a Theory of Mental Representation.” *Mind*, 101 (1992), pp. 243–61.
- Stoljar, D., Forthcoming. “Lewis on Materialism and Experience.” In B. Loewer and J. Schaffer, eds, *Blackwell Companion to David Lewis*.
- , 2010. *Physicalism*. London: Routledge.
- , 2009. “The Argument from Revelation.” In Robert Nola and David Braddon Mitchell, eds, *Conceptual Analysis and Philosophical Naturalism*. Cambridge, MA: MIT Press, pp. 113–38.
- , 2006b. “Actors and Zombies.” In Alex Byrne and Judith Jarvis Thomson, eds, *Content and Modality: Themes from the Philosophy of Robert Stalnaker*. Oxford: OUP, pp. 1–17.
- , 2006a. *Ignorance and Imagination*. New York: OUP.
- , 2005. ‘Physicalism and Phenomenal Concepts’ *Mind and Language*, 20 (5), pp. 469–94.
- Tye, M., 1992. “Naturalism and the Mental.” *Mind*, 101 (1992), pp. 421–41.
- Wilson, T., 2002. *Strangers to Ourselves*. Cambridge, MA: Harvard University Press.
- Wright, C., 2000. “Self-knowledge: The Wittgensteinian Legacy.” In C. Wright, B. Smith, and C. Macdonald, eds, *Knowing Our Own Minds*. Oxford: OUP, pp. 13–46.

24 Personal Identity: Are We Ontological Trash?

Mark Johnston

It has long struck me that there is an unnoticed problem about personal identity. There are semester-long courses on personal identity. (I confess to having taught several myself.) Yet no one would propose to teach a semester-long course on ship identity or snake identity. Is that just due to the characteristic narcissism of our species? There is indeed a vast literature on personal identity, and little written on ship or snake identity, and so if we think courses should cover the relevant literature there is a place for a semester-long course on personal identity, but not on the other two. However, that observation just pushes the bump along the carpet. Why the vast disparity in the sizes of the relevant literatures? Is that just due to the narcissism of our species?

In one sense, I believe that it is. A human being naturally experiences his or her own continued existence as a primary phenomenon, something not simply reducible to an arrangement of matter in space-time; and if one is able to imaginatively enter into one's own coming into being and ceasing to be, those events appear to be fundamental ontological changes, at least in the sense of being not simply reducible to the rearrangement of matter in space-time. Reflection on what one's death consists in leads one to consider our continued existence as the continuation of a consciousness, an arena within which all of one's mental events take place. Absent belief in the next life, death appears as the final end of this consciousness, and so we are led to feel that our continued existence requires the existence of our consciousness. But a consciousness is paradigmatically something which presents itself as relatively self-contained; certainly it does not reveal itself to be simply some arrangement of matter in space-time, even if somehow it actually turns out to be no more than this.

How far this sense of ourselves as primary phenomena is a local cultural phenomenon, having to do with a particular history of religious and philosophical intensification of our own subjective sense of ourselves, is an issue for another time. (Obviously, the extensive discussion of personal identity in Buddhism and particularly Buddhism's focus on overturning the conviction that we are primary phenomena, are highly relevant here.¹) For now, it

is worth noting that John Locke initiated the tradition of serious discussion of personal identity in the West by claiming that neither the persistence of matter, nor the continuation of a particular animal's life, nor even the persistence of a spiritual substance is necessary or sufficient for personal identity. Everything depends, he said, on the continued existence of numerically the same consciousness, where the persistence of a consciousness appears to be a primary phenomenon, even though for Locke it did not consist in the persistence of any material, spiritual, or living substance.²

In a similar vein, Derek Parfit begins his highly influential discussion of personal identity in *Reasons and Persons* with the claim that we are "non-reductionists" about personal identity, that we consider ourselves to be "separately existing entities" distinct from our brains and bodies, whose existence involves a "further fact" whose holding is "all or nothing," "never a matter of degree" and "never a conventional matter."³ While I am broadly sympathetic to the idea that we do have a false picture of our natures as persons, Parfit's initial characterization of our "non-reductionist" outlook has always struck me as subtly mistaken in a way that disables all of his arguments for the claim that personal identity, as opposed to the mental relations which usually accompany it, is not what matters in survival.⁴ After all, Locke arrives at the idea of "same consciousness, same person" by rejecting the importance for our continued existence of souls, spiritual substances and Cartesian egos; and it is clear that for Locke a consciousness is not "separately existing" in Parfit's sense, even though a consciousness's persistence is a further fact, not exhausted by facts about brains or bodies or any other material or spiritual thing. By saying that we regard ourselves as primary phenomena, I mean that we can be naturally led to endorse this further fact view about our continued existence, *however* we concretely fill that out in our imaginations, be it by way of a belief in a soul or in a Lockean consciousness.

Without highlighting the antecedent view that we are, somehow, primary phenomena, the whole topic of personal identity can look picayune. In a recent article, Stephan Blatti gives the following exemplary introduction to the problem of personal identity.

The problem of personal identity emerges from two incontrovertible facts about entities of our kind. First, each of us came into existence at some point in the past; each of us has persisted for some period of time; and each of us will cease to exist at some point in the future. Second, our ontological careers are marked by change. We grow and we diminish; we learn and we forget; we live and we die. The challenge is to account for these two sets of facts in an informative way: in particular, to persuasively articulate the conditions under which we come into, remain in, and ultimately go out of existence.⁵

There is one thing about this characterization that should be obvious; rightly or wrongly, *we* are here being given no special status. Blatti's remarks may be paralleled in the following way:

The problem of ship identity emerges from two incontrovertible facts about entities of that kind. First, each ship came into existence at some point in the past; each ship has persisted for some period of time; and each ship will cease to exist at some point in the future. Second, ships' ontological careers are marked by change. They gain and lose parts over time, they are refurbished and subject to wear and tear, they are built and they rot away. The challenge is to account for these two sets of facts in an informative way: in particular, to persuasively articulate the conditions under which ships come into, remain in, and ultimately go out of existence.

So far, so good. The relation of personal identity is the relation of numerical identity restricted to the case of persons, and the relation of ship identity is the relation of numerical identity restricted to ships. In the first case we need an account of the persistence conditions of persons; that is, an account of what changes they can survive or *be around after*, and in the second case we need an account of the persistence conditions of ships, namely an account of what changes they can survive or *be around after*. Survival in this basic and seemingly unassailable sense is thus logically connected to numerical identity by way of existence: if something x survives some event e then there exists after e a thing y such that $x = y$. This connection holds whether or not we think of persistence as "endurance," as having all of your essence present at each time at which you exist, or instead as "perdurance," as having your essence laid out across time, in the fashion of a temporal parade of items or a cross-time mereological sum of items existing for the briefest of intervals. In the case of perdurance, the question of something surviving a given event is the question of whether numerically one and the same appropriately connected temporal parade or mereological sum of appropriately connected temporal parts exists before and after the event.⁶

However, there are some things that one could be led to think about ships and other artifacts that will seem very disturbing if they turned out to be true of us as well. For example, ships are what we might term "ontological trash." That is, in the nearest spatio-temporal vicinity of anything that has a good claim to be a ship there are other ship-shaped arrangements of matter with different conditions of persistence from the claimant to be the ship. Consider for example, ship-shaped arrangements of matter with slightly more demanding persistence conditions than those of ships, and ship-shaped arrangements of matter with slightly less demanding persistence conditions than those of

ships. These other things, even if we deny each of them the title of being the ship in question, are ontologically on a par with the ship in question. Each is, like the ship in question, a ship-shaped arrangement of matter, where the arrangement has associated with it a certain condition of persistence, a different one in each case. Perhaps by examining our usage of “ship” or, if this is different, our concept of a ship, we can find a limited range of such ship-shaped arrangements of matter which are good candidates to be the ship in question. Even so, the excluded arrangements are not in any way ontologically defective, at least relative to the candidates to be the ship. Indeed there are contexts in which those excluded candidates can become salient, and seem to be the ship in question in the context. This manifests the characteristic sign of ontological trash, namely “*qua*-survival” where we seem to have little option but to suppose that the facts of persistence are exhausted by such remarks as “*qua* or considered as so and so there is here something ship-shaped that survives this event, but *qua* considered as such and such there is not here something ship-shaped which survives this event.”

Ontological trash, like ordinary trash collected in trash heaps, is surrounded by trash, and is not demarcated from the surrounding trash by any special unity condition or principle of self-maintenance over time. Some pieces of trash in the trash heap may become salient to us, but the pieces that do not become salient in one another way are, many of them, on a par with those that do.

Ontological trashiness is to be distinguished from vagueness in the constitution of a material complex; we can adopt a way of dealing with vagueness as semantic indecision over a range of precise material complexes and yet ontologically trashiness will still be with us. The precise material complexes can each be ontologically trashy. To picture the contrast between vagueness in constitution and ontological trash, think of a 15-ball rack of pool balls in American pocket billiards. Whenever there is one sort of rack, say an eight-ball rack with the eight-ball centrally located as per the rules of that game, there is, co-incident with it, another rack suitable for 15-ball rotation, a game in which the eight-ball has no special role, and so can occur at any position in the rack of balls. Here there is no vagueness in the constitution of either rack; each definitely includes the 15 balls and only them. But the two racks—if we reify the racks—have different persistence conditions. The eight-ball rack does not survive swapping the central eight-ball with another ball in the rack, but the coincident 15-ball rotation rack does survive. Despite their distinct conditions of persistence, the two racks are on a par (and on a par with many other racks, which we could define for many other games). They are mere arrangements of balls. If someone swaps the central eight-ball with a ball at a vertex of the rack, and someone else asks “Is that still numerically the same rack?” all the sensible answers are answers given in terms of *qua*-survival, for

example “It is numerically the same qua rotation rack, but not numerically the same qua eight-ball rack (for there is no longer an eight-ball rack there).”

Nor does the idea of ontological trash and associated qua-survival depend on adopting the idiom of coincidence of material complexes with different conditions of existence. A friend of temporal parts could avoid a certain sort of coincidence by insisting that there is just one cross-temporal sum of stages of pool balls; they deserve the title of “the eight-ball rack” because each stage involves a triangle-shaped arrangement of temporal parts of pool balls with a temporal part of an eight-ball in the center. But once again, if we swap balls around and ask “Has the rack survived?” the right thing to say is always something like “It has not survived qua eight-ball rack, but it has survived qua rotation rack.” All we have are cross-time sequences of arrangements of temporal stages of pool balls, along with more or less restrictive conditions on being the same rack. The conditions are all on a par; no one looks more “entity-making” than any other.

That artifacts are ontological trash and accordingly admit only of qua-survival is the more complicated truth behind the often made but literally false remark to the effect that the survival of artifacts is a conventional matter. That is not literally true: when was the convention with respect to ships adopted, and where is the evidence that it is common knowledge among the supposed participants in the convention? When Derek Parfit says that we come to the topic of personal identity believing that our identities are “not a conventional matter and not a matter of degree,” he thereby suggests that we might leave the topic with those convictions overturned by the arguments he offers. That suggestion is plainly false; no argument could show that numerical identity, the relation between a thing and itself, is a matter of degree; and the diagnosis of our visceral anti-reductionism about personal identity is incomplete if it is represented as mere anti-conventionalism about personal identity. Instead we come to the topic of personal identity supposing that we are somehow primary phenomena, where this is inconsistent with our being—even non-conventionally—ontological trash. It is also an outlook at odds with the thought that all the facts about which events we survive involve mere qua-survival.

To put some flesh on these bones, and make more perspicuous the idea of mere qua-survival, consider the old Ship of Theseus case and a modern variant on it.

The Old Case: The ship of Theseus sets out on a long voyage, and its constituent planks are slowly replaced one by one at sea. The process of replacement continues until it is complete. All of the original planks have been shed and are floating in the sea. A canny entrepreneur follows in his own ship picking up all the planks. When he has them all he speeds ahead to the harbor and assembles the planks in just the same arrangement they

were found in at the start of the voyage of the ship of Theseus. The ship which is now made of replacement planks also enters the harbor with Theseus on board. There is now a dispute as to which ship is the real ship of Theseus, the very one that left the harbor with Theseus on it.

Statistically speaking the standard “intuitive response” is that Theseus’s ship survived the slow replacement of planks on the ocean, and so is the ship consisting of entirely different planks from those it had when it left port. But consider this new case.

The New Case: The ship of Theseus is now located in a museum. The head curator has fallen in love with it, and wants to possess it at all costs. He forms a plan. Every night he makes a replica of one of the ship’s planks in his garage at home. He takes it into the museum early the next day and swaps it with the plank it matches. In such a fashion, he collects all the planks from the museum and assembles them into an exactly matching ship in his (huge) garage. The police catch him and accuse him of stealing the ship of Theseus from the museum. The curator’s lawyer makes a devastating observation in court. “Your Honor, this is just the same sort of slow replacement of planks which preserved the identity of the ship of Theseus when it was on the ocean; by the same token then, the ship of Theseus is not in my client’s garage. It is still in the museum, having undergone its second complete replacement of planks.” The judge throws out this line of defense. The curator is convicted of being in possession of the ship, and of having stolen from his museum.

Statistically speaking, the dominant intuition here is that the judge is right, the curator has stolen the ship from the museum. He moved it, by slow replacement of planks, to his garage. But what then is wrong with the defense lawyer’s line of argument?

A natural explanation is that different contexts can make us focus on different processes involving the preservation of ship-shaped arrangements of planks. When we are made to focus on a ship in use at sea, we trace the continuous ship-shaped sequence of changing planks as the continuation of the ship in question, but when we are made to focus on a ship as a historical artifact then we give much more weight to the preservation of its historical constituents.

So we have qua-survival when it comes to ships. As to whether a ship here survives as a result of this process of slow replacement of planks, it depends on whether the ship-qua-ship-in-use is what we have in mind or whether, instead, we have in mind the ship-qua-historical artifact.

In the first case, the ship-qua-ship-in-use follows the line of the continuously ship-shaped thing that crosses the ocean; in the second case the ship-qua-historical-artifact follows the line of the ship-shape thing being constructed from the planks left behind from slow replacement.

This thought “the issue of artifact survival depends on how you look at things” should not be understood as the mad thought that our more specific ways of looking at or conceiving of things *thereby bring other things into being*. Rather the thought should be understood in the following way. Our more specific ways of conceiving of things—say as either the ship-qua-ship-in-use or as the ship-qua-historical artifact—*select* from among things that are already there. That gloss on the thought can be filled out in a variety of ways.

We might suppose that there is at each time a plenitude of coextensive objects, each with a different condition of survival, some of which get teased apart by this or that change. So as “the” ship of Theseus sets out on the ocean, there were many ship-shaped objects that were coincident, and with continuous replacement of planks in one ship and slow side-assembly of another ship, two of these ship-shaped objects came apart. One way of describing the process of continuous replacement of planks in one ship and slow side-assembly of another ship—the way of describing things that we employ in the first case—highlights the career of the ship-qua-ship-in-use. Another way of describing the process of continuous replacement of planks in one ship and slow side-assembly of another ship—the way of describing things that we employ in the second case—highlights the career of the ship-qua-historical-artifact.⁷

Alternatively, we might suppose that there are sequences or parades or cross-time sums of short-lived objects, temporal stages of ships if you will, and in the two cases at hand there are two such divergent sequences or parades or sums. The cases simply make the one and then the other the salient thing to be the ship under consideration.

Either way, ships and other such complex arrangements of matter are ontological trash. Take some kind K of complex arrangement of material parts. In the nearest spatiotemporal vicinity of anything that has a good claim to be a persisting K there are other K-shaped arrangements of matter with different conditions of persistence from the claimant to be the K. These other things, even if we deny them the title of being the K in question, are ontologically on a par with the K in question. Each is, like the K in question, a K-shaped arrangement of matter, where the arrangement has associated with it a certain condition of persistence, a different one in each case. Perhaps by examining our usage of “K” or, if this is different, our concept of a K, we can find a limited range of such K-shaped arrangements of matter that are good candidates to be the K in question. Even so, the excluded arrangements are not in any way ontologically defective, at least relative to the candidates to be the K.

Indeed, as the two cases just discussed show, there may be contexts in which those excluded candidates can become salient and seem to be the K in question in the context. This manifests the characteristic sign of ontological trash, namely “*qua*-survival” where we seem to have little option but to suppose that the facts of persistence are exhausted by such remarks as “*qua* or considered as a K there is nothing that survives this event, but *qua* or considered as something very K-like there is something which survives this event.”

Now here is the solution, not to the problem of personal identity but to the problem of the existence of courses on personal identity. As a matter of psychological fact, we resist thinking of ourselves as ontological trash. We think of ourselves as primary phenomena, and this emerges most clearly when we enter imaginatively into our own deaths, at least if we understand death as final obliteration.⁸ Something fundamental seems gone from the world; indeed it is not too much of an exaggeration to say that a world seems gone. That sense is not captured by claims to the effect that *qua* K you cease to exist, while *qua* K# you continue to exist.

Could I survive Star Trek-style teletransportation? “*Qua* mind or personal-ity, yes; *qua* organism, no.” Whatever that remark amounts to, there is something more that I want to know; namely could I survive *period*, could I simply exist after such an event, or would such an event just simply be my death, understood as an absolute end. In contemplating our deaths, there seems to be a further absolute fact of non-existence for which there is no analog in the case of mere material complexes.

The course on personal identity really begins when we ask what to make of this seeming. What could we possibly be, and what could our persistence conditions over time amount to, if we are not ontological trash? True atoms or simples, that is utterly fundamental particles, are not ontological trash in the sense defined above, but we know we are not such things. So the course needs to maintain an ominous sense that we could turn out to be ontological trash, so that here philosophy may take away one’s death, understood as an absolute event. Of course, in doing so it would also take away one’s life, at least when it is understood as a primary phenomenon.

1 What is the Proper Method in (the) Philosophy (of Personal Identity)?

One virtue of maintaining, throughout the course on personal identity, the ominous sense that we might turn out to be ontological trash is that it serves to remind us that our “intuitions” about personal identity may simply be driven by false beliefs about our natures as persons.

The philosophy of personal identity emerged as a going concern in the 1960s, thanks to the work of such philosophers as David Wiggins, Bernard

Williams, Sydney Shoemaker, John Perry, Derek Parfit, and others inspired by them. Many of these Anglophone philosophers worked explicitly within the idiom of “analytic philosophy” and supposed that the real task of the philosophy of personal identity was to illuminate our *concept* of personal survival by means of organizing our intuitions about survival or continued existence, intuitions gleaned from a wide range of real and imaginary cases.

The fact that the target was a *concept* made the “method of cases” a viable approach. We are highly competent with the concept of personal identity; we have applied it successfully in a wide range of cases throughout human history, and in the common run of cases we have widespread and indisputable knowledge of who is, and *was*, whom. So we must have at least an implicit grasp of the application conditions of the concept of personal identity, and this tacit knowledge of the concept’s application conditions can be manifested equally well in real and imaginary cases. After all, our competence in applying a concept is a capacity to tell whether or not the concept applies, *however* reality might happen to turn out (perhaps within certain degrees of normalcy). Herein lies the evidential role of the imagination in the philosophy of personal identity; imaginary cases and our (supposedly) uniform reaction to them serve as well as real cases in articulating the conditions of application of our concept of personal identity.

If things work out well, the method of collecting and organizing our reactions to cases delivers an “analysis”; that is, an account of a special sort of necessary and sufficient condition or set of conditions for the application of the relevant concept, namely a necessary and sufficient condition or set of conditions that could be recognized as correct simply on the basis of a certain sort of ideal reflection on our tacit understanding of when to apply and when to withhold the concept in question. Therefore, the relevant verdicts and the resultant analysis could be delivered from the armchair, i.e. without any significant empirical investigation; so it was sometimes said that the relevant analyses could be known *a priori*; roughly, in a condition approximating to blissful ignorance of the empirical facts.

In the case of the concept of personal identity the dominant method in analytic philosophy was then to collect “intuitions” about real and imaginary cases of personal survival and ceasing to be, and then bring those intuitions into some sort of reflective equilibrium that bore on the question of the necessary and sufficient conditions for an arbitrary person’s survival. The result would be the filling in of the details of the relation R in an *a priori* (and necessary) bi-conditional of this form:

x, considered at t, is numerically the same person as y, considered at t*, if and only if xRy.

A specification of R would entail a specification of the “identity” or better “persistence” conditions of persons, a specification of what changes they could and could not survive. Thus arose the old analytical question: Is R a matter of x and y having the same body, or being the same organism, or having the same consciousness, or having the same mind (however that mind might be embodied), or having the same separable immaterial soul?

There are many worries that have and can be raised against this whole approach to the question of personal identity, but here is one that has not yet been noted.⁹ Consider the widely believed claim that there are separable souls that leave the bodies they animate at death and begin a journey in the underworld or the overworld, souls that can be re-incarnated in new bodies or—in an alternative version—re-embodied in their resurrected bodies. A large majority of human beings in the last 2,500 years have believed in the separable soul. If this belief were true then it would have a controlling impact on any satisfactory account of the conditions of personal survival. That is, if it were true then it would significantly constrain the filling out of the details of R. But if the proper upshot of the study of personal identity is an *a priori* bi-conditional of the above form, and if the truth or falsehood of the soul hypothesis radically constrains the details of R, then the truth or falsity of the hypothesis that we are separable souls must then be an *a priori* matter.

The problem with this whole approach is that the hypothesis that we are separable souls is manifestly not an *a priori* matter. It is an empirical question that cannot be settled by reflection on the application conditions of our concepts. Whether we are (at least in part) separable souls is an *a posteriori* question that cannot be answered by mere intuition delivered from the armchair.

The relevant empirical facts, not available to us if we simply loll around in the armchair, are these. As we know more and more about the brain, even the highest mental functions seem to have definite brain functions as their condition *sine qua non*. The intriguing specificity of cognitive loss, depending often on the precise location and extent of this or that brain lesion, continually confirms that brain function cannot be overridden as a source of mental capacity. Even in cases of recovery from the specific cognitive losses produced by local brain damage, there is, significantly, no reported phenomenology of memories of an intact thinking soul being “locked inside” an inept, because damaged, brain and body. The thoughts and mental capacities were just not there, it seems. Yet an immaterial bearer of these mental capacities and thoughts need not be damaged just because the brain is damaged.

To develop this line of thought just a little, consider trepanation, one of the oldest surgical procedures, in which holes are made in the skull and through the durus, the tissue surrounding the brain, in order to drain a stroke-induced hematoma that has been putting pressure on the brain, thereby often causing unconsciousness, or at least all the bodily signs of unconsciousness. Many

victims of unconsciousness caused by a subdural hematoma now recover full consciousness as a result of successful trepanning, or “craniotomy” as it is now called.

Why, then, are there no reports of the phenomenology of “being locked in” when the victim “wakes up” after craniotomy?

Contrast the genuine cases of being mentally “locked in” thanks to almost complete bodily paralysis. Such reports arise because in such cases the higher centers of the brain remain intact. But on the substantial dualist conception which has resort to an immaterial soul even these higher centers are mere instruments by which thoughts have their effects. There are, as it were, still “higher” centers, located in the soul or independent spiritual substance that drives the brain, centers that brain damage would leave intact.

Yet temporary brain damage leading to unconsciousness is not, as it happens, phenomenologically like bodily paralysis, contrary to what the hypothesis that our mental life is the life of a separable immaterial soul would predict, at least given natural auxiliary assumptions. Does the blood that leaks out under the durus in subdural hematoma, also somehow leak into the soul?

This kind of point goes beyond the mere correlations revealed by lesion work and MRIs. In the face of all the bodily signs of unconsciousness found in severe cases of subdural hematoma, the natural dualist expectation would be that if and when the pressure on the brain was relieved and the patient awoke, he or she would report having been “locked in” to a bodily prison. This is just what we do not find.

I do not say that there are no alternative auxiliary hypotheses that would make the facts concerning the phenomenology of recovery from subdural hematoma *consistent* with the hypothesis of the separable soul. I just say that any friend of this hypothesis should be surprised and slightly dismayed by such facts.

On the other side of the ledger, there are the startling reports, more or less spontaneously produced in roughly 10 to 15 percent of resuscitated cardiac arrest cases, of what are now called “out-of-body” experiences. The experiences, which seem to their subjects to be happening during a period that coincides with the clinical death of their own bodies, involve such things as the sense of leaving one’s body and of looking down at the medical personnel pumping one’s chest in their attempts at resuscitation. As the experience develops, one seems to be traveling through a tunnel to a bliss-inducing white light. As one moves in the light, dead relatives, and even old pets, are encountered, and in some cases one is presented with a review of one’s life. This, of course, would represent the beginning of an empirical vindication of the soul hypothesis.

Here is what I take to be least controversial in this controversial area. There is a genuine phenomenon that goes under the name of the “out-of-body

experience.” However, in investigating this phenomenon, one does not find *robust* evidence of distinctive knowledge of the external world that could only be gleaned from the ostensible vantage points of the supposedly disembodied subject. If one has left one’s body and is looking down upon it, then one could be expected to take in facts about, say, the emergency room that are not available to the normal viewers, there on the floor. Experiments have indeed been proposed, even partly performed, but what we do not have is a decisive case that clearly passes the obvious test. I have in mind a cartoon that effectively presents the obvious test: we see an emergency room in which a cardiac arrest patient is being resuscitated by doctors. Mounted high up on the back wall of the emergency room is a sign whose message is visible *only* from near the ceiling of the room.



What if such obvious tests were frequently passed by the resuscitated patients? What if they reported reading such signs as “You’re Dead!” “Eat at Joe’s!” “Medicare Won’t Be Covering This!” or whatever happens to be displayed at the moment of their deaths. And suppose we could rule out collusion and suggestion and remote cognition from the hospital bed? What then?

I must confess that that would be enough for me. We should then have to seriously consider the idea of locating our mental lives in independent substances, substances whose mental functioning can outlive the functioning of their associated brains. So far, no such disturbing reports have been collected in any well-controlled setting; but they might turn up or they might not turn up. (My bet is that they won't, but that is based on other empirical evidence of the sort cited earlier.) All of this just serves to highlight what should be anyway be obvious, namely that the separated survival of the soul is an empirical question. It remains an empirical question even if, as many think, it is an entirely settled empirical question.

The point is that this is the death knell of the old "analytic" approach to the problem of personal identity. For it shows that the correct filling out of R in the bi-conditional

x, considered at t, is numerically the same person as y, considered at t* if and only if xRy.

is an a posteriori or empirical matter. And this means that it is not accessible simply by way of articulating our tacit understanding of the application conditions of the concept of being the same person. Our intuitions about cases of personal survival, be those cases real and imaginary, are just beliefs we have, and they may well be false, especially in the *recherché* cases designed precisely to tease apart the very things—the persistence of the body, mental continuity, the persistence of the individual personality—which go together in the central core of ordinary cases that dominate our experience of the continued existence of persons.

There is a natural weakening of the target bi-conditional that might seem responsive to this objection from the a posteriori or empirical status of the proposition that we are (at least in part) separable immaterial souls. Given the ideology behind the method of cases, the best the conceptual analyst can hope for is a series of a priori conditionals, conditions underwritten by our implicit conceptual knowledge; for example:

If persons are (at least in part) immaterial souls then Q is the relation that has to hold between x and y for x considered at t to be numerically the same person as y considered at t'.

If persons are not (even in part) immaterial souls then S is the relation that has to hold between x and y for x considered at t to be numerically the same person as y considered at t'.

So now, the response goes, we can see that the method of cases can proceed as intended, as long as we understand our initial target relations as

various—including say Q and S—with our ultimate choice among them resting on the empirical evidence of the sort already cited.

The original objection now takes its definitive form when we ask: just how various are the initial target relations left open by what we know a priori, that is to say in the armchair? Notice that there seems to be no specific relation S such that the following bi-conditional is both non-trivial and knowable a priori:

If persons are not (even in part) immaterial souls then S is the relation that has to hold between x and y for x considered at t to be numerically the same person as y considered at t'.

For, clearly, there are various distinct conditions of identity for various kinds of “soul-free” persons. If we are always minded animals, and could not cease to be such without ceasing to be, then one relation—one close to that specified by the bodily criterion—will be relevant. However, if we are always somehow-or-other embodied minds and could not cease to be such without ceasing to be, then another relation—one close to that specified by the psychological criterion—will be relevant. So are we minded animals or are we embodied minds? The deliverances of the method of cases give out precisely here; hence the unresolved dispute between what was called the bodily and the psychological criteria of personal identity. Anyone who claims to have resolved that dispute by the method of cases will simply find “differing intuitions” on the other side. So here too we must resort to empirical means, we must use all we collectively know and all of our capacities for argumentative ingenuity to settle the question.

What is coming into clear view here is that we really wanted to know the answer to this question: what is it that we are? Are we (at least in part) always, and must we always continue to be (at least in part), immaterial souls? Are we always, and must we always continue to be, minded animals? Are we always, and must we always continue to be, embodied consciousnesses? Or are we something else entirely? The method of cases, properly thought through, is impotent to decide between such questions. The most it can do is explore some fairly direct consequences of one or another of these hypotheses, when they are assumed to be true. But we wanted to know *which* hypothesis is correct.

Why did we philosophers overestimate the power of the method of cases, understood as a way of articulating our tacit “conceptual” knowledge? Why did we think we could resolve hard cases based on the articulation of the application conditions of our concepts? The question bears centrally on the status of analytical philosophy understood as the analysis of concepts.

One suggestion, due to Sarah-Jane Leslie, is that there is a tendency on the part of philosophers to misrepresent “conceptual knowledge” as tacit

knowledge of universally valid necessary and sufficient conditions; that is, tacit knowledge of universally quantified bi-conditionals that hold in all possible worlds. As against this, there is a growing body of empirical evidence that application conditions of our concepts do not take the form of exceptionless necessary and sufficient conditions, but rather rely on generic connections among concepts, connections that admit of exceptions *that are nonetheless not counterexamples*.¹⁰

Some philosophers might say that each concept by its nature either applies to a given situation or does not apply, so that if our conceptual knowledge is exhausted by such generic connections then the right thing to say is that we are employing many concepts of personal identity, namely all of those which accord with the verdicts entailed by the relevant generics, but differ over the remaining cases. Psychologists do not talk this way; but there lies no great sin. They can clearheadedly say that the proper word for what those philosophers are calling “concepts” is instead “property.” Given this translation manual, one can hold to two very natural theses; there is *an* ordinary concept of a persisting person, and yet it underdetermines exactly which property we have in mind.

So it may be that according to our ordinary concept of a persisting person the generic belief to the effect that persons survive if their individual minds continue on, and the generic belief to the effect that persons will survive if their bodies are kept alive and functioning, are both true. Notice that a negative verdict on our prospects of surviving teletransportation does not yield a counterexample to the first generic, but only an unusual exception. It can seem to be a counterexample because we sometimes misconstrue generic connections as universal quantifications, perhaps with restricted antecedents. While a white raven is a counterexample to “All Ravens are black” it is not a counterexample to “Ravens are black.” Male kangaroos are counterexamples to “All normal Kangaroos carry their young around in pouches” but they are not counterexamples to “Kangaroos carry their young around in pouches.”

Accordingly, the opinion that a person does not survive in a persistent vegetative condition is not at odds with the generic belief that persons survive if their bodies are kept alive and functioning, because that opinion concerns an exceptional case, which only became very salient with the rise of advanced life-support technology. Generics allow for just such exceptions, and so if our “a priori or conceptual knowledge” of the conditions of personal survival is exhausted by such generics then the analytic method of cases with its focus on the persistent vegetative condition, teletransportation and the like was just a mistaken attempt to exploit our conceptual competence precisely where it delivered no verdict, namely the very cases which that competence simply ignored.

That would predict that even an ideal and exhaustive application of the method of cases would leave the “hard cases” unresolved. Is this not what we actually find when we consider teletransportation and the like? The bodily criterion and the pure or “wide” psychological criterion differ over whether the very same organism or brain has to survive in order for the very same person to survive, and no amount of clever marshalling of intuitions seems to decide this issue.

Indeed, in the massive core of cases of ordinary survival from day to day, many sources of evidence for personal survival, such as persistent bodily integrity and mental continuity, converge and agree, whereas the whole philosophical charm and supposed utility of the imagined cases in the literature on personal identity lies precisely in teasing these elements apart. The obvious question arises: might we not have thereby *undermined* our ability to make good judgments about personal identity when considering these very cases?

Moreover, when we take the trouble to look, we do not find much evidence that in tracking objects and persons through time we are actually deploying knowledge of *sufficient* conditions for cross-time identity. Instead, as a matter of empirical fact, it appears that nature saves us cognitive labor by having us “offload” the question of sufficiency onto the objects and people themselves—if I may put it that way. The idea of offloading can be put in terms of the motto “I don’t know what the (non-trivial) sufficient conditions for identity over time are, but I do know a persisting object when I see one.” Objects of various kinds are salient to us, they attract our attention, and we track them through space and time. Those objects either survive or fail to survive, as an objective matter of fact, determined by their respective natures. As long as they do not manifest changes in respects we know to be important for their kind, we are ready to credit them as having survived, even if we remain properly agnostic about what their persistence actually consists in. So we should be prepared to discover that in tracking objects we are deploying knowledge of *some* necessary conditions for their survival over time—the thing can’t explode into smithereens, for example—but not of any non-trivial sufficient condition. The objects just take care of themselves in this regard, they either persist or cease to be; to *witness* such outcomes we need not know any sufficient condition for their persistence. Accordingly our perceptual system can save us cognitive labor; in disclosing persisting rather than momentary objects to us, it allows us to offload what would otherwise be the cognitive task of stitching together momentary objects into persisting wholes by way of necessary and sufficient conditions for being included in the persisting whole.

We “offload” then onto a class of objects when our basic way of tracing or re-identifying them is *criterionless*, in the sense of not employing sufficient conditions for cross-time identity so as to move from neutral evidence to a conclusion concerning the identity over time or persistence of objects in the

class. Instead, the persisting objects capture our attention at various times and over time.

All of this suggests that reliance on the analytic method of cases was misguided, and in at least two ways. First, in so far as we are interested in the structure of our *concepts*, we need to look hard into actual cognitive science, and particularly into the detailed questions of the development of our conceptual competence and its bearing on tracing objects through change. Once we do that, we will have less confidence that even our sheer “conceptual” competence can properly be studied in the armchair. Second, the whole interest of the question of personal identity, as it appears say in Locke, lies in the promise of an account of *what we are, and of what changes we can in principle survive*. (Locke, after all, was centrally concerned about the resurrection; about how it was in principle possible for the dead to reappear on the “Great Day.”) Our concepts of a person and of numerically the same person may indeed function to pick out some kind of thing and some kind of relation; but the real interest lies in the nature of the kind of thing in question and the nature of the relation in question.

We must turn our attention to kinds and relations and away from the concepts that serve to highlight them. After all, our concepts can serve to highlight things that they systematically misrepresent, so that an explication of our concepts will then be a rehearsal of error. Again, in the case of personal identity, there is a natural suspicion that our conviction that we are primary phenomena—either by being immaterial souls or *somehow or other*—lies at the core of our self-conception and so probably infects our concept of personal survival.

Those thoughts re-enforce the idea that the proper philosophical method, here as elsewhere, is not to limit oneself to the impoverished realm of conceptual or a priori knowledge, knowledge somehow deriving from, or embedded in, our competence with concepts (or alternatively knowledge deriving from our competence with meaningful terms of our language). The proper method is to use *all* one knows and all one can find out, in the most ingenious ways one can. Philosophy is integrative theoretical vision combined with argumentative ingenuity, deployed at a fairly abstract level. Philosophy has no special province; but so far from marginalizing philosophy, this liberates it. On the other hand, it follows that we philosophers are now under a clear obligation to learn a lot more science than the analysts of old deemed relevant. We need to get out of the armchair and look into things.

Nonetheless we have to begin where we are, and it seems that we do view ourselves as primary phenomena. So it has to be part of any reasonable methodology to determine how far such a view coherently hangs together with the rest of what we know. On the one hand, there is the evidence, ever so briefly glossed above, to the effect that we do not find in ourselves extra units, such

as immaterial souls, that could not be ultimately constituted out of the basic items found in a materialist ontology; on the other hand, there is the antecedent conviction that we are primary phenomena, and so not mere arrangements of matter in space-time.

Could it then be that we are *special* sorts of arrangements of matter in space-time, arrangements that are not ontologically trashy, that do not overlap with other such arrangements on a par with us? What could that idea amount to?

2 Do We Have the Persistence Conditions of Organisms?

In *Particulars and Persistence* (1984) I attempted to find, within a materialist ontology, items that are not ontologically trashy, thereby suggesting that our convictions about our distinctive ontological status might be saved to some degree. Such an ontology includes material simples, material complexes ultimately made up of simples, and more substantial complexes, namely organisms, in which the constituents are taken up in a life. Although we resist thinking of ourselves as mere material complexes, overlapping with a massive number of similar material complexes with differing persistence conditions, where these are all on a par, would we offer the same resistance to the idea that we are essentially organisms of a certain kind, namely animals?

It depends of course on the characterization of organisms and on how far that distinguishes them from *mere* material complexes like artifacts and rocks and stellar dust. Roughly, the idea was that an organism, unlike a mere material complex has something close to what Spinoza called a “conatus,” namely a dynamically operating power of self-maintenance which works to keep the organism in existence in its natural environment.

Though the material constitution of an organism might be vague, the hope was that nothing in the *nearest* vicinity of an organism has the same material constituents taken up in another dynamically operating power of self-maintenance. The hope depends on a particular way of spelling out just what an organism is.

Organisms are marked off from other material things by their distinctive mode of being. And they are marked off from other living things, like organs or cells, by their distinctive mode of being *alive*. Organisms have the capacity for dynamic self-maintenance of their life functions. These life functions are dispositions to do various things, and when there are a number of such dispositions, we say that we have a living organism.

In the case of a human organism, those life functions include locomotion to sources of food, ingestion, metabolizing of nutrients, breathing, oxygenation of the blood, and excretion. So matter is absorbed into the organism, exchanged, and expelled. Hence the dynamic material character of an organism, a feature

it shares with organs like livers and brains. But organs and livers are not self-maintaining in the way organisms are. By this I mean that when it comes to the organism, the dynamic exchange of matter with its environment happens within limits, which preserve a continuing basis for those very life functions. The life functions of the organism thereby maintain themselves—they are self-maintaining—because they maintain a physical basis for the very dispositions that they are. Organisms, as long as they remain alive, have the power to dynamically maintain a physical basis for the very dispositions which are constitutive of their life. Death is the loss of that power.

The brain may also be said to have life functions as well; sensing is a life function, a function that helps maintain the life of the animal, and more complex mental processes may help maintain the life of the animals that exhibit them. The brain is such that some of its operations—brain events—constitute sensing and thinking, the liver breaks down toxic substances in the blood, and so on. But neither of these organs has the capacity to dynamically maintain within itself a basis for these characteristic life functions. When the organism dies at somatic death, the organs then soon die, precisely because they do not have this capacity for dynamic self-maintenance.

It is important to see that “alive” is polysemous; it is one thing for a brain or a liver to be alive and another thing for an organism to be alive. A brain, as opposed to some of the cells that make it up, is alive only if it is still capable of *carrying out* its characteristic life-function, namely providing the basis for thought; similarly, *mutatis mutandis*, for a liver. An organism is alive if it is still capable of dynamic self-maintenance of its life functions. These are quite different necessary conditions on “being alive.” All your organs could be intact, and so yet to die, after you are killed, say by a heart-stopping electric shock.

That point is underwritten by biology’s distinction between “somatic death,” the death of an organic body or organism, the death of cells, and organ death, the death of the organism’s constituent organs. Somatic death is a global physiological phenomenon; the signs of somatic death are the cessation of heartbeat, breathing, movement, and brain activity. Cell death is a smaller-scale biochemical process in which various cells in various organs lose the capacity to subserve life-functions. Cell death sets in some time after somatic death, and with it comes organ death. That is why the organs of the dead can be usefully transplanted into the living if those organs are recovered quickly enough. Cell death in these organs has not progressed to the point where the organs have lost the capacity to subserve various life-functions. As it happens, cell death sets in at different times for different types of cells. Left to themselves, brain cells may survive for about 10 minutes after somatic death, while those of the heart can survive for about 20 minutes, and those of the liver for about 30 to 40 minutes. (Some estimates of these average times are more liberal, others more conservative.)¹¹

Since a brain dies in one way, and an organism dies in another way, a brain is not an organism. That is the other side of the fact that since a brain is alive in one way and an organism is alive in another way, a brain is not an organism.

Even if a brain or a severed head or a cerebrum are kept on life support, they *never had the capacity for dynamic self-maintenance of any of their life-functions*, including mentation, the very life-function they subserved. This is why we should not say that severed heads, and separated brains and cerebrums, are organisms or animals. It is at odds with what biology teaches about the difference between somatic death and organ death, and this is because it is at odds with what biology teaches about the difference between somatic life and organ life.

An organism, given its pattern of dynamic self-maintenance, has a claim to be a distinguished entity that stands out against the variety of mere material complexes that surround it. It is not in general true that wherever we have an organism, we also have in the nearest vicinity of that organism something else that is dynamically self-maintaining but in a variant way. Organisms therefore do not seem to be ontological trash, on a par with mere material complexes. Is this the best approximation for our idea of ourselves to be found in a purely material world?¹² That consideration in favor of the view that we are organisms is to be distinguished from sheer intuitions to the effect that we would survive all and only what an organism would survive. Indeed, as we shall see below, our intuitions are actually at odds with such a claim—see the “brain-transplanting” case.

Another consideration in favor of the view that we are organisms, likewise independent of appeal to the method of cases, goes by way of the fact that much of our knowledge of personal identity derives from visually tracing persons in a fashion which is not too different from our tracing of dogs, of cats, and of deer. That is, we arrive at easy and offhand knowledge of personal identity by observing bodily continuity and the absence of radical changes in behavior. This looks very much like the visual tracing over time of dogs, cats, and deer. And *they* surely are animals, that is, organisms of a certain kind.

Just to recall these easy and offhand ways in which we have knowledge of personal identity: I seem to have rich knowledge of a host of humdrum facts of personal identity over time as I observe my students in class, and yet, often enough when they are not speaking, I bear no more complicated epistemic relationship to them that I do to the deer in a my backyard. That is, I trace them over time by way of carefree observation of bodily continuity, sometimes augmented by observing some sort of consistency in bodily behavior.

Still, I have massive knowledge of my students’ identities over time without, it seems, relying on auxiliary hypotheses concerning regular connections between the kind of things I observe, namely animals of a certain sort, and metaphysically more exotic entities such as souls, let alone bare subjects of

experience who could survive any amount of bodily and psychological discontinuity. During the lulls in class discussion when the students fidget, or daydream, or look to their teacher and try simply to take in what he is saying, I am still in a position to notice persisting animals of a certain sort and thereby come to have knowledge that, for example, the person before me in *that seat* is numerically one and the same person that I saw just a minute or so earlier.

It seems that if I go by any rule in arriving at such knowledge, it is the simple (generic) rule: same human organism, same person.

Taking this fact seriously makes it very difficult to maintain that what we ultimately turn out to be are *bare* subjects of experience; that is, simple beings whose continued existence is not constrained either by gross bodily or mental continuity. Another thought about how it might be that we are indeed primary phenomena is thereby quashed.

Even if bare subjects of experience were to turn out to exist, they are not the things that we are tracing when we trace persons. Otherwise, our offhand methods would hardly deliver massive knowledge of facts of personal identity. Relative to our easy and offhand methods of tracing persons, judgments about the persistence of bare subjects would be highly adventurous conjectures. Nevertheless, we do have rich, if humdrum, knowledge of personal identity over time, at least in the massive core of ordinary cases. Indeed, it is among the most secure sorts of knowledge that we possess.

Taken together, the consideration to the effect that the idea that we are organisms is the closest realizer of the idea that we are primary phenomena (at least within a materialistic framework), and the consideration that in tracing ourselves in the easy an offhand ways we do, we are relying on the same methods we use to trace other organisms, in particular animals, suggest that the place to start in discussions of personal identity is with the view that we are organisms of a certain sort.

3 The Brain-transplanting Intuition

The view that we are essentially organisms, along with the logically weaker view that we can survive all and only what an organism can survive,¹³ are at odds with the most famous appeal to the method of cases in the philosophy of personal identity. Here is Sydney Shoemaker's discussion drawn from his book *Self-Knowledge and Self-Identity*.

It is now possible to transplant certain organs . . . in such a way that the organ continues to function in its new setting. . . . [I]t is at least conceivable . . . that a human body could continue to function normally if its brain were replaced by one taken from another human body. . . . Two

men, a Mr. Brown and a Mr. Robinson, had been operated on for brain tumors, and brain extractions had been performed on both of them. At the end of the operations, however, the assistant inadvertently put Brown's brain in Robinson's head, and Robinson's brain in Brown's head. One of these men immediately dies, but the other, the one with Robinson's head and Brown's brain, eventually regains consciousness. Let us call the latter "Brownson." . . . When asked his name he automatically replies "Brown." He recognizes Brown's wife and family . . . and is able to describe in detail events in Brown's life. . . . Of Robinson's past life he evidences no knowledge at all.¹⁴

As a matter of statistical fact the dominant response to this case is that Brown is numerically the same person as Brownson; that we in effect go where our living brains go. The intuition seems all the more compelling if we add that Brown's brain was kept alive and functioning during the operation, so that there was continuous consciousness subserved by that brain while it was in transit from Brown's body to Robinson's body. Brown goes where his brain goes, according to the widespread and strong intuition. If we treat this intuition as evidence, we can then reason this way. Brownson is not the same organism or human animal as Brown. In fact, Brown's de-brained body—let's call it "Brownless"—could be provided with enough in the way of brain-stem tissue, transplanted from still another source, so as to be kept alive in a persistent vegetative condition. It can seem that this human "vegetable" Brownless is identical with the very *animal* that once exhausted Brown's bodily nature. Brownless is that animal in a surgically mutilated condition.

Brown has not only ceased to be the animal organism that he was, but during the transition, he simply consisted of a brain. But a brain is an organ, not an organism. No brain is an organism. Some philosophers have denied this, but in doing so they have forgotten the different ways in which organism and brains are alive. Even if a brain is kept on life support, it *never had the capacity for dynamic self-maintenance of any of its life functions*. A brain's being alive is just its having enough functioning cells to perform its characteristic life-function, namely mentation. So Brown does not survive only what an organism is capable of surviving. He does not have the persistence conditions of an organism.

Notice, however, that this argument turns on a classic appeal to an intuition about an odd case, a case which is *precisely designed* to tease apart mental continuity (holding between Brown and Brownson) and the organic physical continuity (holding between Brown and Brownless) that secures the survival of a human organism. As argued earlier, we have good reason to reject the sheer appeal to such cases as probative, on the obvious ground that in separating out mental and physical continuity the presentation of such cases may well undermine the precondition of our being good judges of personal identity,

namely the convergence of these two forms of continuity, which in everyday life we take to be the reliable signs of personal identity.

Moreover, even if we grant that something like the principle that persons survive if their minds survive does lie behind the application conditions of our concept of being the same person, say in the sense that we are guided by this generic in evaluating the ordinary cases of personal identity over time, that is in the massive range of cases where mental and physical continuity line up, then it would still be fallacious to reason the following way about the brain transplanting case:

- (a) Persons survive if their minds survive.
- (b) Brown's mind survives when only his brain is kept alive and functioning.
- (c) So, Brown survives when only his brain is kept alive and functioning.
- (d) A brain is not an organism.
- (e) So Brown can survive something an organism cannot survive.

The second premise can seem quite plausible, given that the supervenience of mind on brain functioning is widely accepted. In the ordinary case, the continued existence of a mind is secured by the continued functioning (of the relevant sort) of the associated brain, and it is natural to suppose that this would continue to be so even if the brain was kept alive and functioning (in the relevant way) by advanced medical procedures.

However, even granting all that, the conclusion does not follow, since the first premise is a generic, plausibly understood as governing what characteristically happens in the ordinary run of things. Unfortunately for the proposed argument, the case at hand is precisely designed to fall outside the ordinary run of things, since it separates out mental and physical continuity.

4 Remnant Persons

There is however a problem for the view that we have the persistence conditions of organisms, one which I take to be very considerable. None of the proposed ways out seems successful.¹⁵

Here is a conviction that many will share, one which organizes some of our thinking about persons and physical reality, in particular what we now know about the neural basis of consciousness: you can't bring a person into being simply by removing tissue from something, and then destroying that tissue, unless that tissue was functioning to suppress mental life or the capacity for mental life. A developing fetus might have a massive tumor in its developing brain, which suppresses its mental life, and perhaps even its capacity for mental life. Given that, we can understand how removing the tumor could allow a

person in Locke's sense—a thinking reflective being that can consider itself as itself at various times and places—to be present for the first time. But you do not bring a person into existence by removing an arm, or a leg, or even a sustaining torso. A person's coming into being is not that kind of extrinsic matter. Nor is this a merely generic truth; rather it holds universally.

This principle

(No creation) You don't cause a person to come into being by removing tissue, unless that tissue is suppressing the capacity for reflective mental life.

organizes a good deal of empirically driven thinking about the effects of medical interventions, and in particular what we now know about the way in which our having a mental life of the sort sufficient for personhood depends just on our neural functioning.

The principle is thus not a mere upshot of the armchair application of the method of cases. It is more like a detailed denial of certain sorts of occult effects; namely, the creation at a distance of the kind of mentality sufficient for personhood without any effect on the underlying neural functioning.

Now it is relatively easy to see that the result that we do not have the persistence conditions of organisms follows from the no-creation principle in conjunction with a plausible supervenience principle to the effect that if a brain is artificially kept alive and fully functioning, complex mental life of the sort sufficient for personhood will be maintained, even if that brain comes to be no longer surrounded by a supportive organic body. The supervenience principle implies that in such cases there will be a "remnant person," a person who is obviously not an organism because he is not constituted by a living organic body of the appropriate sort. The no-creation principle then implies that the remnant person did not just come into existence, but was there beforehand. But the only person there beforehand was the human organism from which the live brain was taken. So the person who was a human organism survives, that is continues to exist, in a condition in which he or she is not an organism.

That is to say, a person that was constituted by an organism can cease to be an organism—his constitution can be reduced to that of a severed head or an envatted brain—without ceasing to be. So persons do not have the persistence conditions of organisms.

Once again, this last point is crucially not to be established by an "intuition" about Shoemaker's imaginary case, but by abstract reflection about the consequences of what we know about causation and supervenience.

Let's use an imaginary case, not to pump intuitions, but to illustrate just how the argument works.

5 Really Gruesome Guillotining

In the next reign of terror the guillotine returns, but in an even more gruesome form. The aristocrats, otherwise known as the “one percenters,” are placed face down with their heads leaning over a platform. A huge metal block falls from 12 feet above the platform, and completely obliterates the victim’s body from the head down; the head flies forward, and is caught by an official who quickly attaches it to a medical device which keeps it alive and functioning. The crowds execrate the head for the next few days, until it dies off.

We are to suppose that a hapless aristocrat, who was the victim of this horror, was an organism; that is, he had an organic body of the relevant sort. This body was obliterated at guillotining, so that the organism that the aristocrat was made up of ceased to exist at just that point. We are also to suppose that the aristocrat’s head is kept alive in such a way that the brain of the aristocrat still functions as an organ of mentation; that is, some of its operations are or constitute thoughts, indeed reflective thoughts involving profound despair about the future and the like.

Of course, as things now stand keeping a severed head alive for any significant amount of time may not be medically possible, but it is definitely not metaphysically impossible. Indeed, actual guillotining might well have included the terminal experience of the world spinning around, as one’s head rolled into the basket. The brain takes a while to die, so that actual guillotining may not have brought mental life to an immediate end.

So we can legitimately suppose that really gruesome guillotining would leave behind something—something made up of a living head, artificially kept alive—with the active capacity for reflective mental life, something which anticipates a bleak future and frets about this, and so something which is a person.

The supposition that there is still a person around after really gruesome guillotining can be supported by appeal to a certain sort of supervenience claim. While the aristocrat was intact, it was his brain’s functioning in a certain sort of way that secured the kind of mentality sufficient for personhood; when only the aristocrat’s head survives, that brain might still function in the relevant way for some time, so that it continues to secure the kind of mentality sufficient for personhood. In accepting that, we need not be assuming that it is numerically one and the same person before and after really gruesome guillotining, but only that there is a person—a remnant person—around after really gruesome guillotining, one wholly constituted by a head, or as it might be in even more gruesome variants by a brain, or by a cerebrum.

So there are then two possibilities: either the remnant person is numerically the same person as the original aristocrat, or he is not. Suppose that the remnant person is a numerically different person from the aristocrat. Then

either he was there all along, a distinct person from the aristocrat located just where the aristocrat's head was located, or he came into being as a result of really gruesome guillotining. But the only relevant person there beforehand was the aristocrat, whose head, brain, and cerebrum all survived the gruesome event. So it follows that the remnant person was brought into being by really gruesome guillotining. Simply as a result of the (violent) removal of tissue, tissue which was not in any way suppressing a capacity for reflective mental life (but instead helping to sustain that capacity) a person was caused to come into existence.

However, this conclusion is at odds with:

(No creation) You don't cause a person to come into being by removing tissue, unless that tissue is suppressing the capacity for reflective mental life.

So the thing to conclude—the only thing that does not lead to this or that repugnant consequence—is that the aristocrat continues to exist after really gruesome guillotining; he comes to be (wholly constituted by) a head kept alive and functioning. In the even more gruesome versions, he comes to be (wholly constituted by) a brain kept alive and functioning or by a cerebrum kept alive and functioning. It follows that in such cases a person ceases to be an organism without thereby ceasing to be. Obviously, that conclusion is also at odds with what is now called “Animalism,” the view that our persistence conditions are just those of animals. For something is an animal only if and while it is an organism of a certain sort. In really gruesome guillotining, a person ceases to be an animal without thereby ceasing to be.¹⁶

6 Are We Embodied Minds?

The view that we have the persistence conditions of organisms might have made us ontologically distinctive; but in so far as one is led to the view that one is a primary phenomenon by reflection on the absolute character of our deaths, understood as the final end of consciousness, any purely body-based account of our persistence will not seem to get to the heart of the matter.

We have already reviewed the evidence, drawn from subdural hematoma and the like, that we are not separable souls or spiritual substances; even so, this leaves untouched a variety of views on which the persistence of one's mind is crucial to one's identity. We could be minds that are essentially embodied in this or that way.

In contemporary analytic philosophy the most well-known version of the view that we are embodied minds is called “Neo-Lockeanism” or “The Wide Psychological View” a position ably defended by Anthony Quinton, Sydney

Shoemaker, and David Lewis.¹⁷ These thinkers took Locke to be offering a clearly inadequate memory criterion for personal identity over time, and they therefore sought to bolster Locke's position by adding other mental connections such as the persistence of beliefs and desires, the tendency to carry out intentions just formed, the immediate carry-over of character, and so on, so as to include all the direct mental links that are typically found from one moment to the next in a single conscious life. Neo-Lockeanism thus arrives at the position that a person at one time is identical with a later person just in case there is a chain of such direct mental links *uniquely connecting* the first mentioned person and the second. So according to Neo-Lockeanism you will exist at a later time just in case you have a unique, and sufficiently close, mental continuer at that time. (Lewis found an ingenious way to drop the requirement of uniqueness.¹⁸)

To go into the view in a little more detail, this so-called Neo-Lockean position is a version of *psychological reductionism*, since it has it that truths about personal identity have as necessary and sufficient conditions statements about the holding of relations of mental continuity and connectedness. Connectedness involves the holding of direct psychological connections, such as the persistence of beliefs and desires, the connection between an intention and the later act in which the intention is carried out, and the connection between an experience and a memory of that experience. Importantly for what follows, connectedness can come in twice over in the statement of the conditions on personal identity. All psychological reductionists require that, if two person states are states of the same person, then psychological continuity, the ancestral of strong or predominant psychological connectedness, holds between them. Some psychological reductionists also require that no two such stages be entirely unconnected psychologically.

So-called Neo-Lockeanism is more informatively styled *wide* psychological reductionism because it has it that mental continuity and connectedness can constitute personal identity even if the holding of these relations is not secured by its normal cause, the persistence of a particular human brain. Any causal mechanism that operates so that these psychological relations hold will do. The identity over time of any particular human body or brain plays no strictly indispensable role in the identity of a particular person over time. Any particular human body or brain is just one causal means among others for the holding of the relations of psychological continuity and connectedness that themselves constitute a particular person's survival. So the characteristic upshot of the wide psychological view is the consequence that a person could survive teletransportation. And, of course, wide psychological reductionism, by disavowing the importance of the persistence of the normal cause of mental continuity and connectedness, also implies that one could survive really gruesome guillotining.

On the other hand, the so-called *narrow* psychological reductionist holds that we only have the persistence of a person when mental continuity and connectedness is secured by its normal cause, the persistence of a particular human brain. So while it implies that we could survive really gruesome guillotining, narrow psychological reductionism construes teletransportation not as survival but as replacement by a duplicate.

Both forms of psychological reductionism will have the consequence that a person does not survive in a persistent vegetative condition, along with the related consequence that no person was an embryo devoid of psychological states. For there is nothing mental in the persistent vegetative condition for a relation of psychological continuity to hook on to; and the same goes for the as-yet mindless embryo.

As noted earlier, the appeal to “intuitions” here is less than probative; the method of cases is at best a (not very telling) indication of the structure of our concepts, whereas we are here concerned with our own natures. The crucial thing to see is that in that in either of its forms, psychological reductionism counts us as ontological trash! On either variant of the view, in the near spatiotemporal vicinity of any persisting person there will be any number of overlapping very-person like things, each with a different temporal extent, and hence each with a different precise condition of persistence. This arises from the fact that mental connectedness is a matter of degree.

To see this, suppose our psychological reductionist tells us that some degree *D* of mental connectedness is required over the short intervals that make up a *person's* life and that some (much lesser) degree of mental connectedness *D** is required over the entirety of a person's life. The justification of these imposed degrees does not come down from a metaphysical Mt Sinai; it can only consist in a specification of the degrees of connectedness found for the most part in the typical mental lives exhibited by persons. But now consider both some slightly more and some slightly less demanding variants on *D* and *D**. Many pairs of these variants will specify the mental lives of very person-like things, things which are partly spatiotemporally coincident with the persons whose mental lives conform to *D* and *D**. These very person-like things are ontologically on a par with persons, just as the very ship-like things, with cross-time unity conditions that are slightly more or slightly less demanding than the cross-time unity condition for ships, are ontologically on a par with ships.

Of course, the very person-like things won't, according to psychological reductionism, deserve the appellation “person,” but some of them are very close to being persons. How close they are depends just on which types of person-like things we consider. For example, according to psychological reductionism, there is to be found in the life of any person who lives 70 years or so, the person-like thing, that is the psychologically continuous and connected

thing, which corresponds to the last 10 years of that person's life. That thing will exhibit a certain degree and style of psychological continuity and connectedness, and it is hard to see how it can be that only one degree and style of psychological continuity and connectedness, namely the one pertaining to persons, is entity-making. The point obviously generalizes, so that in the life of any person who lives N years, there is to be found the person-like things which correspond to the last M years (where M can take any value less than N) of the person's life. There are thus a host of such "latter day selves" to be found within the life of any given person; certainly one comes into being on each successive birthday, and equally certainly one comes into being on each passing day of the person's life.

So as well as a given person, psychological reductionism had better recognize all these "latter-day selves" that are not persons, but somewhat person-like. (The alternative is just special pleading to the effect that, somehow, only the cross-temporal unity condition associated with persons is entity-making). Once again, this is just a way of highlighting, within one theory of our natures as persons, the fact that we are ontological trash. The crucial question is whether the latter-day selves are sufficiently person-like to deserve the respect owed to persons. The obvious problem is that if they are in this sense sufficiently person-like then our ordinary practices of blame and punishment will look too much like *collective* blame and *collective* punishment, and will be morally objectionable for the same obvious reason, namely that they impose pain and other psychic costs on beings who did not commit the offences in question. Blaming and punishing the person who committed the crime seems to make sense, but it can only be done by imposing pain and other psychic costs on a host of latter-day selves that are innocent of the crime, since they came into existence only after the crime was committed.

I suppose the thing to say is that these latter-day selves are not only not persons but are not even very person-like. That is at least a tenable position; for the latter-day selves, unlike ordinary persons, are not born, they do not mature and grow into personhood, but rather just appear with a given mental life on a given day. Suddenly coming into being in this way, may seem to make a thing very unlike a typical, paradigm person. A friend of psychological reductionism could thus avoid the worry that our ordinary practices of blame and punishment are as wrong as collective blame and punishment, by emphasizing that these latter-day selves, though somewhat person-like are not *very* person-like. So the requirements of respect for persons do not naturally expand to encompass latter-day selves.

It is much harder to make a case of that sort for what we might call "earlier selves." Indeed these earlier selves, as we shall see, could be intrinsically just like things which would be, according to psychological reductionism, bona fide persons.

Consider the case of Dum and Dee, two identical twins raised in symmetric environments, where they are monitored and modified regularly in order to keep their brain functioning and hence their mental life as close as possible. Dum ends up living six months longer than Dee, months during which Dum is engaged entirely in reveries that are fairly psychologically disconnected from Dum's earlier mental life. By hypothesis, however, there is still one person Dum and one mental life which he has; for despite the reverie there is at least a degree D^* of mental connectedness exhibited over his whole mental life. The same is true of Dee's mental life, there is at least a degree D^* of mental connectedness exhibited over his whole mental life. (In fact, thanks to the absence of six months of reverie, there is a higher degree of mental connectedness exhibited over Dee's whole mental life.)

Notice now, that as well as Dum and Dee, there is another very person-like thing – call him Don – who consists of the embodied mental life of Dum up to but not including the six months of reverie which completed Dum's mental life. According to psychological reductionism, Don does not deserve the appellation "person." This is because there is more mental life (the six months of reverie) continuous and connected with Don's, but which is not included in Don's mental life. In the idiom of temporal stages, Don is not a *maximal* sum of psychologically connected and continuous stages. But Don is very, very person-like. One way to see this is to notice that he is intrinsically just like Dee, Dum's twin who died six months earlier than Dum! Throughout their lives Don and Dum have been in exactly the same types of physical and "inner" mental states, in exactly the same sequence. Even though "Don" does not, according to psychological reductionism, deserve the appellation "person," there is nonetheless a person that Don could not be more similar to, namely Dee.

Once this point is clear, we don't really need twins like Dum and Dee to make the point. Given psychological reductionism, we have an ontological trash heap; within any person's life there will be found the lives of many, many person-like things that are intrinsically just like possible persons. That is, for each of these other things there could be a typical person that is intrinsically just like it.

Notice that this is a consequence just of the continuity and connectedness requirements of psychological reductionism; the mental element distinctive of psychological reductionism is not really playing a crucial role in the argument! Suppose we instead consider an account of personal identity in terms of bodily continuity and connectedness, with $D^\#$ and D^\wedge as the relevant degrees of overall connectedness. The same sort of considerations will lead us to recognize a similar ontological trash heap; within any person's life there will be found the lives of many, many person-like things that are intrinsically just like possible persons.

Now we are in a position to see that there are really two versions of the view that we are organisms; the one is just a bodily continuity analysis of personal identity, which treats us as ontological trash, while the other purports to find a unique and distinctive self-maintaining disposition which is there so long as the body is alive. Whether there really is room for this second sort of view within a naturalistic framework depends on delicate questions about the nature of dispositions. Are they genuine powers distinct from their categorical bases or is it simply that we have a dispositional idiom, which functions as a useful way of picking out a range of categorical bases that simply mediate similar causes onto similar effects? On the second view, we can capture all the facts by describing the categorical bases; the dispositional idiom is just a useful shorthand. It would then follow that there is nothing ontologically distinctive about self-maintaining powers, and so nothing ontologically distinctive about life or organisms.

7 Why Is It So Bad to Be Trash?

Consider the person, Dum, and the very person-like thing, Don, who is almost, but for Dum's last six months, coincident with Dum in space-time. Don is intrinsically like Dee, a person on anyone's view. Dee certainly deserves the respect owed to persons; his interests cannot be overridden just because *someone else* might benefit. For example, Dee cannot simply be required to undergo a tedious amount of study to deliver a benefit for me; if we have come to an reasonable agreement with appropriate considerations involved, then it may be legitimate for me to use him as my research assistant; but unless there is something like this in play I am just using Dee, that is failing to respect his interests as a person, and this is so even if I do not here a peep of protest out of him. What I am doing would then be wrong; moreover my intending to implement a plan which involves treating him in this way is to that extent wrong.

Consider another case. I go to the doctor and he informs me that I am the site of a bizarre form of parthenogenesis. Lodged alongside my right lung there is a fully formed small person, who is miraculously still alive and cannot be removed without killing both of us. Tests show that he is not in any evident distress there, except when I go in for vigorous exercise, such as running on the treadmill. The test show that as my lungs rapidly contract and expand, this small person goes into a kind of writing agony, although I only notice this as a slight twinge in my chest. Even though I hear no peep from him, I come to believe that he is genuinely suffering without adequate recompense every time I train intensely. So I must give up my intention to train for the decathlon. (Obviously, this really is an imaginary case.) Continuing to train like that causes him too much suffering, and he has a legitimate interest in not suffering

without adequate compensation. Even my *intention* to train vigorously is now shown to be wrong. I am obliged not to intend that. The legitimate interests of other persons place serious constraints on what we can do and on what we can intend to do.

But now consider this “No prejudice” principle:

The legitimate interests of very person-like things—things which are such that there could be a typical person intrinsically exactly like them—deserve respect, and thereby place serious constraints on what we can do and on what we can intend to do.

I do not really see how this principle could be seriously denied. If one thought that persons were primary phenomena, so that there was nothing in their nearest vicinity very person-like, then one might see the principle as in effect equivalent to the principle that persons deserve respect, and so as adding no further obligations to our already existing roster of obligations to persons. But once one sees that in the nearest vicinity of any person there are indefinitely many person-like things—indeed so person-like that they could wholly duplicate a person—it then seems arbitrary to restrict one’s respect only to those duplicates of persons (real or possible), which also deserve the appellation “person.” For deserving that appellation is a matter of being a salient piece of ontological trash in an ontological trash heap in which many, many things are ontologically on a par. In fact, on the continuity theories we have discussed, those duplicates of real or possible persons which nonetheless do not deserve the appellation “person” fail to deserve this appellation only because of what happens *after* they cease to exist. (Recall, they are followed by bodily or mental stages that are continuous with what has gone before.) Once we see that this is the only basis for denying then the appellation “person” (at least according to one or another continuity theory) then it does seem arbitrary to respect persons and not respect things that duplicate would-be typical persons.

We have seen that continuity theories entail that we are ontological trash, so that we are found in ontological trash heaps along with many, many, very person-like things. It will then seem arbitrary not to accept the “no prejudice” principle. The legitimate interests of very person-like things do seem to constrain what we can legitimately do and what we can legitimately intend.

8 Must We Then Be Feckless?

A repugnant consequence follows. There is a characteristic form of apparently rational deliberation, intention, and discipline that dominates the mental lives of adults, at least those who are not in a state of extreme *gelassenheit* or “letting

things be." It will follow that this form of deliberation, intention, and discipline is morally wrong. If some continuity theory of personal identity theory is true we must no longer live or think that way, otherwise we are in relevant respects just like the imagined me when I get on the treadmill indifferent to the discomfort of the little person lodged alongside my right lung.

This form of deliberation, intention, and discipline might be called "prudential time-sacrificing." Here is an example, showing how humdrum and everyday this orientation is. I have decided to go to Italy at the end of summer. Things go so much better in Italy if you have a little Italian. My Italian is very rusty, and I can only get it up to speed by an intensive couple of weeks of work. I find language work very tedious, there is for me a particular kind of mental pain that accompanies it; nonetheless it is in my interest to be able to speak some Italian when I am in Italy. Being a (marginally) self-disciplined person, I sacrifice time, effort, and a certain amount of real discomfort for the anticipated future benefit. Not to do this would seem to be *feckless*, that is it would seem to be a failure to give the right weight to my future interests.

That whole orientation is morally wrong, it now seems. For consider many of the very person-like things in my spatiotemporal vicinity who I can know in advance will undergo the uncomfortable grind of brushing up on Italian. I will be imposing the uncomfortable grind on them by imposing it on myself. But I also know that many of them will cease to be before I get to Italy. So I am imposing a cost on many, many very person-like things where I know that there will be no compensating benefit. The no-prejudice principle tells me that this is just wrong.

I would guess that most of us engage daily in such prudential time-sacrificing deliberation, then form the relevant intentions and act on them. If some continuity theory of personal identity is true this whole form of mental life is a moral error. It is *just as if* we were thereby flouting the legitimate interests of other people, by imposing on them costs where there was no compensating benefit. The right thing to be is utterly feckless!

Would a thoroughgoing utilitarian picture of our obligations, both to persons and the very person-like, somehow save the day? Many very person-like things undergo the mental pain of working on Italian and many very person-like things end up there in Italy enjoying easy communication with the locals. But many very person-like things who undergo the mental pain of working on Italian do not get to Italy and enjoy easy communication with the locals.

This last group represents a significant and as-yet unnoticed tax on overall utility that must be paid off by the overall utility enjoyed by those who do get to Italy. Such (so far unnoticed) taxes are likely to render morally wrong many forms of what were thought to be apparently innocent prudential time-sacrificing behavior. The upshot is that we should be much more feckless than our mentors have advised; we should not be moved by the small or medium

advantages that present psychic pain and sacrifice might secure in the future. We should consider the previously unnoticed side-effects on the very person-like things that cohabit our world with us. We should only make present psychic sacrifices when the anticipated benefit is large enough to pay off the tax, and leaves still more utility on the table.

Imagine a class of very unhealthy students preparing months ahead of time for a long summer trip to Italy. They start learning Italian in the winter, but then in the spring a deadly virus kills off many of the students before they can get to Italy. What looked like a good scheme from the utilitarian viewpoint is thus rendered decidedly non-felicitic, and hence objectively wrong if there are enough deaths in the interim. For those who died endured the pain of language work but not the benefit of easy communication in Italy. If enough die this tips the whole scheme into the territory of negative utility. Even from the utilitarian point of view, it is only morally defensible to plan such trips if the survivors are likely to have a truly wonderful time in Italy.

As the collective case suggests, the real cost of prudential time sacrifice has gone unnoticed thanks to the fact that we think of ourselves as primary phenomena, distinguished from the many things that share our bodily and mental lives. Either we should adopt a deontological view of our duties to others and be utterly feckless, or we should adopt a utilitarian view of our duties to others and be considerably more feckless (unless of course we are already quite feckless).

I do not seriously intend this as a justification for fecklessness, even though I believe that it is hard to resist the idea that some defense of fecklessness follows from the view that we are ontological trash, and hence from any continuity theory of personal identity. If we are ontological trash, time-sacrificing prudence always involves hitherto unnoticed costs imposed on very person-like things, including things which fail to deserve the appellation "person" only because of what happens *after* they cease to exist.

The real upshot, I believe, is that we are not made to think of ourselves as ontological trash. Our basic pattern of self-regarding deliberation is seriously undermined if we are ontological trash. However, it is very difficult, at least within a materialist ontology, to see how we could be anything other than ontological trash.

Now, at least, the course on personal identity is well and truly underway.

Notes

- 1 For an overview, see Mark Siderits 2007.
- 2 See "Of Identity and Difference" in Locke's *Essay Concerning Human Understanding*, Book 2, Chapter 27, Section 9. My description of Locke depends on the new interpretation of Locke as significantly different from the fabled "memory-theorist" of

personal identity. For this new interpretation, see Galen Strawson 2012 and Shelly Weinberg forthcoming . The interpretation is anticipated in the discussion of the idea of an arena of presence in my *Surviving Death* (2010). I argue that there are no consciousnesses or minds in “Why there are no Visual Fields (and no Minds Either)” 2011, pp. 231–42.

- 3 See the beginning of part three of his *Reasons and Persons* (OUP, 1984).
- 4 In my “Human Concerns without Superlative Selves” (2003) Parfit’s arguments for his central claim that personal identity is not what matters in survival are shown to depend upon what I call the argument from below. Parfit discusses this criticism in a number of places, most noticeably in his prize essay at www.ammonius.com. As I point out in the early part of the fifth chapter of my *Surviving Death* Parfit there offers a plausible account of when the argument from below is a good one, but even on his own account it is not a good one in the case of personal identity. This disables his main line of argument to his central conclusion in part three of *Reasons and Persons*.
- 5 This appears at the opening of his paper “Animalism” (2006).
- 6 This is the account of endurance and perdurance which I now prefer, not the account taken up from my dissertation by David Lewis in his *Plurality of Worlds* (1986), where endurance is defined as being wholly present over time. Suppose you gain or lose parts over time, you could still endure, but what does it mean to say the whole of you is present from moment to moment if your parts change over time? In the case of an enduring organism, in the sense discussed below, the essence of the organism, namely that there be some organic matter whose structure sustains the self-maintaining life functions characteristic of the organism, is a present condition of the organism at each time at which it exists. (At least if we follow Aristotle’s view that a dead organism is no longer the same thing as the organism that previously lived.)
- 7 See Stephen Yablo 1987, pp. 293–314, Mark Johnston 2006, and Sarah-Jane Leslie 2011, pp. 277–96.
- 8 For a brilliant discussion of entering imaginatively into our own death and appreciating it as an absolute absence, see Valberg 2007.
- 9 For some others, see Kathleen Wilkes 1988 along with my “Human Beings,” (1987) and *Surviving Death* (2010), pp. 44–7. A certain amount of what follows occurs with more detail added in chapter 2 of *Surviving Death*.
- 10 Leslie forthcoming with OUP.
- 11 It is important to remember that capacities or dispositions can be had when they are not operating, and when their manifestations are being temporarily suppressed by other factors; so an organism on life support, say a breathing tube, may not have lost the capacity to breathe; that capacity may simply be suppressed by fluid in the lungs. But when the organism has ceased to have functioning lungs, it has lost one of its life-functions. When it loses all of these it is definitely dead.
- 12 Once again, notice that this point about organisms not being ontological trash is compatible with vagueness in our concept of an atom or a molecule being taken up in the self-maintaining process that is the life of the organism. Nor does the problem of the many, properly treated, show that organisms are ontologically trashy. On this see my “Constitution Is Not Identity” (1992).
- 13 The first view, but not the second, implies that I could not have come into being as anything but an organism.
- 14 *Self-Knowledge and Self-Identity* (1963), pp. 23–4.
- 15 For an extended discussion of the problem, and of the proposed ways out, see my “Remnant Persons: Animalism Undone” (2013).
- 16 For detailed discussions of animalism see Ayers 1991; Carter 1988, 1989, 1999; Mackie 1999; Olson, 1997, 2003; Snowdon 1990, 1991, 1995; Wollheim 1984.

- 17 See Lewis 1976; Quinton 1962, pp. 393–409; Shoemaker, with Swinburne, 1984. Shoemaker clearly had some sympathy for the view that bodily continuity is constitutive of personal identity when he wrote *Self-knowledge and Self-identity* (1963), although he allowed that in exceptional cases the bodily criterion of personal identity could be overridden by the memory criterion. He decisively abandons the bodily criterion in “Persons and Their Pasts,” (1970), pp. 269–85.
- 18 See his “Survival and Identity,” op. cit.

Bibliography

- Ayers, M., 1991. *Locke*. 2 vols. London: Routledge.
- Blatti, S., 2006. “Animalism.” In A. C. Grayling, A. Pyle, and N. Goulder, eds, *Continuum Encyclopedia of British Philosophy*, London: Continuum, pp. 108–9.
- Carter, W., 1999. “Will I Be a Dead Person?” *Philosophy and Phenomenological Research*, 59, pp. 167–71.
- , 1989. “How to Change Your Mind.” *Canadian Journal of Philosophy*, 19, pp. 1–14.
- , 1988. “Our Bodies, Our Selves.” *Australasian Journal of Philosophy*, 66, pp. 308–19.
- Johnston, M., 2013. “Remnant Persons: Animalism Undone.” In S. Blatti and P. Snowdon, eds, *Essays on Animalism*. Oxford: OUP.
- , 2011. “Why There Are No Visual Fields (and No Minds Either).” *Analytic Philosophy*, 4, pp. 231–42.
- , 2010. *Surviving Death*. Princeton: Princeton University Press.
- , 2006. “Hylomorphism.” *Journal of Philosophy*, 103, 652–98.
- , 2003. “Human Concerns without Superlative Selves.” In R. Martin and J. Barresi, eds, *Personal Identity*. Oxford: Blackwell, pp. 260–91.
- , 1992. “Constitution Is Not Identity.” *Mind*, 101, pp. 89–106.
- , 1987. “Human Beings.” *Journal of Philosophy*, 84, pp. 59–83.
- Leslie, S. J., Forthcoming. *Generics and Generalization*. Oxford: OUP.
- , 2011. “Essence, Plenitude, and Paradox.” *Philosophical Perspectives*, 25, pp. 277–96.
- Lewis, D., 1986. *Plurality of Worlds*. Oxford: OUP.
- , 1976. “Survival and Identity.” In A. O. Rorty, ed., *The Identities of Persons*. Berkeley: California University Press, pp. 17–40.
- Mackie, D., 1999. “Animalism vs Lockeanism: No Contest.” *Philosophical Quarterly*, 49, pp. 83–90.
- Olson, E., 2003. “An Argument for Animalism.” In J. Barresi and R. Martin, eds, *Personal Identity*. Oxford: Blackwell, pp. 318–34.
- , 1997. *The Human Animal: Personal Identity without Psychology*. Oxford: OUP.
- Parfit, D., 1984. *Reasons and Persons*. Oxford: OUP.
- Quinton, A., 1962. “The Soul.” *The Journal of Philosophy*, LIX, 15 (July 19), pp. 393–409.
- Shoemaker, S., 1970. “Persons and Their Pasts.” *American Philosophical Quarterly*, VII, 4 (October), pp. 269–85.
- , 1963. *Self-knowledge and Self-identity*. Ithaca, NY: Cornell University Press.
- Shoemaker, S. and Swinburne, R., 1984. *Personal Identity*. Oxford: Blackwell.
- Siderits, M., 2007. *Personal Identity and Buddhist Philosophy*. Farnham: Ashgate.

- Snowdon, P., 1995. "Persons, Animals, and Bodies." In J. L. Bermúdez, A. Marcel, and N. Eilan, eds, *The Body and the Self*. Cambridge, MA: MIT Press, pp. 71–86.
- , 1991. "Personal Identity and Brain Transplants." In D. Cockburn, ed., *Human Beings*. Cambridge: CUP, pp. 109–26.
- , 1990. "Persons, Animals, and Ourselves." In C. Gill, ed., *The Person and the Human Mind*. Oxford: OUP, pp. 83–107.
- Strawson, G., 2012. *Locke on Personal Identity: Consciousness and Concernment*. Princeton: Princeton University Press.
- Valberg, J. J., 2007. *Dream, Death and Self*. Princeton: Princeton University Press.
- Weinberg, S., Forthcoming. "The Metaphysical Fact of Consciousness in Locke's Theory of Personal Identity." *Journal of the History of Philosophy*.
- Wilkes, K., 1988. *Real People*. Oxford: OUP.
- Wollheim, R., 1984. *The Thread of Life*. Cambridge: CUP.
- Yablo, S., 1987. "Identity, Essence, and Indiscernibility." *Journal of Philosophy*, 84, pp. 293–314.

25 Free Will

Ferenc Huoranszki

The expression “free will” is translated from the Latin term “*liberum arbitrium*” which means free choice or free decision. This suggests that freedom of the will is a problem about how choices can be free. And indeed, this is one way to understand the problem about free will. However, as we shall see, there is an equally important but broader interpretation of the problem of free will that directly concerns our actions, and not only the ability to choose, or the ways in which we may exert that ability. It seems to us that we make choices when we face alternatives. But how exactly should such alternatives be understood? Does it make sense to talk about free will even if no alternatives are open to us? Such questions shall preoccupy us in the large part of this chapter, but before we address them we need to say more about the broader philosophical context in which the problem of free will becomes particularly interesting.

1 Free Will and Responsibility

The significance of the question about choice can be best understood if we see how it is related to the philosophical problem of agency. Even if freedom of the will is a technical notion in philosophy and the question about the nature of choice might sound somewhat esoteric, there is an important aspect of our life the significance of which cannot be denied and to which the problem of free will is tightly connected: the problem of responsibility. Although it might be unclear how the notion of “freedom” can be applied to choices, no one would deny that it is applicable to some of our actions. And at least one reason to be interested in the meaning of freedom of action is that free agency is a condition of responsibility.

We hold each other, as well as ourselves, morally or prudentially responsible for certain things we do or omit. But we also think that agents are not necessarily responsible for everything they do (let alone for everything they omit). For instance, sometimes we are not responsible for our actions because we do not do them intentionally. It seems then that the intentionality of our

behavior is often — although not without exception — necessary for our responsibility. However, the intentionality of behavior is not sufficient for responsibility. We would not hold animals responsible even if many of them can also act intentionally to the extent that they can control their own behavior in view of some purpose. A dog can see a cat and chase it, and in doing so he moves his own body with an aim. But this capacity is not sufficient for holding dogs morally responsible for their behavior.

Thus, we can say that whatever free will is, it is that aspect of human agency which explains how we can be morally or prudentially, and not only causally, responsible for our behavior. Understood in this way, freedom of the will is at least a condition of responsibility. It needs to be noted, however, that not every philosopher would regard free will as a condition of responsibility. Some philosophers think that the conditions in which we can have free will are much more stringent than the conditions in which we can be held morally responsible, and hence freedom of the will, though perhaps sufficient, is not necessary for moral responsibility (Frankfurt 1988; Clarke 2003), as the following shows.

In everyday life, we routinely hold each other responsible, and this might be a useful social practice. This practice is not arbitrary in the sense that there are more or less well identifiable conditions that justify responsibility ascriptions. But according to some philosophers in order to justify such practices we need not assume that agents have free will. Rather, free will is the condition of “real origination”: it is a condition of considering ourselves as creators of our own self or a condition of regarding ourselves as the author of our own life. In this sense, free will is the capacity of *self-determination*. It is contentious, however, whether the possession of this capacity is necessary for our day-to-day practice of holding each other responsible.

Why should we distinguish conditions of self-determination from the ordinary notion of agency that is necessary for ascribing responsibility? After all, one can argue that free will as the capacity of self-determination is also a condition of moral responsibility as ordinarily understood. The problem is that it is questionable whether ordinary agents can have the capacity of self-determination. The stoics, for instance, thought that only the wise man can appropriately determine his own state, but then said that it is doubtful whether any man is actually wise.

The philosophical debate here centers on the issue whether the concept of free will can or should be independent of our ordinary practice of ascribing responsibility. Many philosophers think that though it is a contentious issue whether we can have free will in the sense that we can determine our own state, this should not affect our concept of free will as a condition of our responsibility; while others believe that in order to have free will we should be able to determine our self, and if such self-determination is impossible,

then there is a sense in which none of us is responsible (Strawson 1994). In what follows I shall assume that free will is a condition of our responsibility for actions and omissions, without taking side in the issue of origination and self-determination.

2 Alternative Possibilities

As we have seen, the capacity of intentional control is not sufficient for holding someone responsible. What other conditions need to be added? Perhaps the dog is not responsible because he lacks certain *rational and moral capacities*. It would be vain to try to explain to him that it is wrong to chase the cat or to want to cause harm to her. Thus, responsibility requires the possession of some rational capacities which enable agents to make judgments about what is wrong and what is right. However, it is a further question whether the possession of rational capacities is sufficient for responsibility. It seems that in order to hold persons responsible we also have to assume that they are able to control their behavior in the light of their rational judgments. In this sense, agents have free will only if they can exercise a kind of control over their own behavior that is appropriate for their responsibility.

One traditional way to understand this kind of control is to say that agents are responsible only if they can do otherwise, that is if they have alternative possibilities open to them. Two of the most important questions about the metaphysics of free will concern exactly this requirement. First, there is a question about how to understand the nature of these possibilities. Second, some philosophers have challenged the claim that the control necessary for moral responsibility requires the agent's ability to do otherwise. We shall address first the problem of alternative possibilities and control, and come back later to the arguments for the irrelevance of alternative possibilities.

If one can do otherwise, it must be possible for her that she avoids doing something that she does; or alternatively, that she does something that in fact she fails to do. How can we understand such possibilities? Does the existence of free will depend on the satisfaction of certain conditions that might not be satisfied in our universe? Can it turn out that we are not responsible agents after all because in the absence of these conditions free will is impossible? But in which conditions do we have reason to believe that we cannot do otherwise? Obviously, in conditions in which nothing else can happen than what actually happens. And it might seem that in a universe where every event is *determined* to happen nothing else *can* happen than what actually happens. If this is right then determinism and the ability to act otherwise are incompatible.

However, the incompatibility of determinism and alternative possibilities is not an obvious matter. There are many different ways in which we

can understand determinism, and not every form of determinism makes it metaphysically impossible that nothing else can happen than what actually happens. Thus, in order to understand better the problem of free will and determinism, we must clearly distinguish different forms of determinism.

First, we have to distinguish natural determinism from “logical determinism.” “Logical determinism” can also be called fatalism, even if, confusingly, some traditional views of fatalism (like the stoic doctrine of fate) are much closer to natural than to “logical” determinism, and it will be discussed and explained in some detail in the next section. Anyhow, the difference between logical and natural determinism is that natural determinism may or may not obtain in a universe depending on which kind of laws govern the evolution of events, whereas logical determinism, if true, is true in any metaphysically possible universe.

Importantly, however, natural determinism itself can come in different varieties. We need to distinguish physical determinism from social or biological determinism. The latter is related to the ways in which social circumstances and biologically inherited properties can affect persons’ behavior. The former is a view about the system of physical laws that govern the evolution of events in the whole universe. The different forms of determinism are differently related to the possibility of free will as the ability to do otherwise. The truth of logical determinism would certainly be incompatible with the ability to act otherwise. The situation with different kinds of natural determinism is more complex.

It might naturally be thought that the truth of social and psychological determinism are incompatible with the ability to act otherwise, because in those cases the determinism is operating on the personal level. The relation between physical (i.e. on the level of basic physics) determinism and the ability to act otherwise is a lot more difficult issue. *Incompatibilists* argue that physical determinism is incompatible with free will because it is incompatible with the ability to do otherwise. *Compatibilists* about free will deny the relevance of physical determinism to the question of the free will, for the determinism at the basic physical level is essentially sub-personal and so might be thought not to threaten persons’ ability to act or to refrain from acting as they want.

Before I proceed it needs to be mentioned that there is a fourth historically important form of determinism, which has certain affinities with both the logical and the physical forms of determinism: theological determinism. Since knowledge entails truth, if God knows, as he does, all of my future actions, there is a sense in which those actions must occur. Whether or not that sense is compatible with my ability to avoid doing what I do is the question of theological determinism. The issue about theological determinism is similar to fatalism in the sense that it arises no matter what laws obtain in a world. However, in an important respect, the issue of theological determinism

is more similar to the issue of physical determinism than to that of logical determinism. If logical determinism is true, then free will cannot exist. But the interesting philosophical question about physical determinism is whether free will can exist even in deterministic universes. In this respect, the question about theological determinism is similar to the issue about physical determinism: whether free will can exist even if God exists and he foreknows every human action (Craig 1987). For reasons of space, important as it is, we cannot discuss theological determinism here.

3 Fatalism

Logical determinism, or fatalism, is the view that whatever happens is unavoidable no matter what. Fatalists—in this technical sense—claim that if a proposition is true, it is true forever. Otherwise put, propositions do not change their truth value in time. Certainly, the proposition that “E happens” must be true after any time of E’s occurrence. But, fatalists argue, it must also be true before E happens. Of course, we know much more about the past than about the future, since we have records about the past, but we do not have records about what is going to happen. But this is insignificant. After all, there are many past events that we do not know, but this does not entail that there is no truth about them. Similarly, the fact that we do not know the future does not mean that there is no truth about it. But if there is truth about the future, which true propositions represent, and such propositions have their truth value eternally then whatever happens must happen. As a consequence no one can do something else than what she does.

Some philosophers believe that we can only reject fatalism if we deny that propositions cannot change their truth value (Cahn 1967). They admit that *if* a proposition is already true, it cannot change from true to false. But some propositions—propositions about the future contingents—are neither true nor false now. It is in this sense that the “future is open,” whereas the past is not. This proposal is not ad hoc. For instance, it can be supported by certain considerations about the metaphysics of time. It might be the case that we cannot understand temporal asymmetries—like the openness future as opposed to the past—without admitting that propositions about future contingents are neither true nor false in the present.

Hence to say that some propositions about the future are neither true nor false sounds intuitive. However, accepting this leads to certain logical problems, and it is also questionable how this suggestion can answer the problem of fatalism. First, even if “E will happen” is neither true nor false now, it must be true even now that E will happen or it won’t happen. But how can *that* be true if neither “E will happen” nor “E will not happen” is true? After all, a

disjunction is true if at least one of its disjuncts is true. But, according to this solution, neither “E will happen” nor “E won’t happen” is true. More seriously, it seems that the fatalist argument can be restated even if we grant that propositions about the future are neither true nor false in the present, since it must be true of every proposition now that it *will become true* (or become false) in the future. But then it is true now that they will become true (or that they will become false) in the future and hence it is inevitable that the events that will make them true (or false) are going to happen.

So do we have to accept fatalism? Not necessarily. We might say that instead of fiddling with the nature of propositions, we should deny that their “eternal truth” has much to do with the possibility of certain events. After all, even if it is true that certain events cannot happen unless the corresponding propositions are true, this does not mean that the truth of such propositions can make events (including human actions, which are our concern here) inevitable. Even if certain propositions about the future couldn’t be true unless certain events happen, and they are in fact true, it is not the truth of the propositions that makes it the case that those events occur. Thus, even if there are true propositions about the future, this does not imply that the facts that make them true are inevitable and hence it is impossible for us to act otherwise in the future. Logical determinism may be false, and hence it cannot make free will impossible, not because propositions about the future lack truth value in the present, but because their truth depends on what we may or may not do, while what we do does not depend on their truth (Huoranszki 2002).

4 Determinism and the Consequence Argument

There are different arguments for the incompatibility of physical determinism and free will, some based on some pre-theoretical belief about incompatibility, others on rather technical reasoning. In this section we shall discuss the most important argument for incompatibility, the so-called consequence argument. There is another influential argument for incompatibility, called the manipulation argument, which we shall briefly mention later. The consequence argument supports incompatibilism only indirectly by trying to show that physical determinism is incompatible with the ability to do otherwise. The scope of the manipulation argument is broader in the sense that it aims to prove that free will is incompatible with determinism even if free will does not require the ability to do otherwise.

Although we often talk about natural determinism indiscriminately, physical determinism on the one hand, and social and biological determinism on the other, raise different issues. Even if the universe is not physically deterministic, social and biological determinism might be true. And social and

biological determinism can be false even if physical determinism is true. More importantly, the issue of social and psychological determinism sets a different challenge for the reality of free will than physical determinism does. As we shall see, philosophers disagree on whether physical determinism is compatible with agents' ability to do otherwise. It is hard to see, however, how social or biological determinism can be compatible with such ability. Since everyone agrees that the circumstances in which we were brought up, or our genetic heritage—or perhaps both together—determine some of our behavioral dispositions, the real question about social and biological determinism is whether such dispositions themselves are sufficient to explain in every particular situation how we act. According to social and/or biological determinism every piece of intentional behavior must be traced back to such factors, and this is exactly to say that, given their social and biological circumstances, agents cannot act otherwise in any specific situation.

Moreover, it seems that social and biological determinism makes sense only if there are physical possibilities. If what agents do is already physically necessitated there is no reason to assume that their behavior is socially or biologically determined. Physical necessitation *screens off* the relevance of social or biological determination of behavior. If the truth of physical determinism is sufficient to show that agents cannot do otherwise, then the question of social or biological determinism becomes irrelevant. On the other hand, such forms of determinism are obviously incompatible with the ability to act otherwise, since determinism in this context means exactly that the possession of certain socially and biologically determined dispositions is sufficient for acting in the way we do, provided it is physically possible to act so.

Thus, the problem of physical determinism must be clearly distinguished from the issue of social and biological determination. Physical determinism is not a question about how we are "physically predisposed" to act. Rather it is a question about how the nature of physical laws is related to our abilities to act. It is in this context that the question of incompatibility becomes particularly interesting. Incompatibilists claim that if the universe is deterministic—that is if the system of laws that govern the evolution of physical events constitutes a deterministic system—it is impossible for anyone to have the power or ability to do otherwise. Many current debates about the metaphysics of free will consider exactly this issue.

The most important argument for the impossibility to do otherwise in a deterministic universe is the consequence argument. Informally, the argument can be summarized as follows. It is not up to us how the universe was in the remote past. It is also not up to us which laws govern the evolution of physical events. But if determinism is true, then our future actions are the consequences of the remote past and the laws of nature. Thus, it is not up to us what

we do next. The consequence argument is meant to express our firm initial belief about the incompatibility of physical determinism and free will as the ability to do otherwise (van Inwagen, 1983).

However, it seems that some of our other convictions work in the opposite direction. What does exactly “up to us” mean here? Sometimes it means that I can intentionally control my own behavior. Normally, I’m able to intentionally control the movements of my arm, but I’m not able so to control my blood pressure. In this sense, my blood pressure is never “up to me,” but sometimes the movement of my arm is. Apply now the consequence argument. I’m not able to intentionally control the past and the physical laws. If determinism is true then every movement of my arm is the consequence of the remote past and the laws of nature. Does it follow that I’m never able to intentionally control the movement of my arms unless determinism is false? This argument would “prove” that intentional control of actual behavior requires the falsity of physical determinism. But this is implausible, no matter whether or not we are compatibilists.

Incompatibilists respond that they have a different ability in mind: they want to show something about *what* can happen, even if in fact it does not happen. But the ability to control our actions intentionally tells us only *how* we do what we actually do, and not whether we can act *differently*. Certainly, it is not true that if someone is unable to make it that φ , and φ entails ψ , then he/she is also unable to make it that ψ . But according to the incompatibilists, it is still true that if someone is unable to make it that *not* φ , and φ entails ψ , then he/she is also unable to make it that *not* ψ . But why agents’ abilities with regard to what they actually do and their abilities with regard to what they do not do, but can do, should be different given that they are equally powerless with regard to the past and the laws in both cases? In what ways does their inability to affect the past and the laws deprive agents of their present abilities, whether exercised or not?

Incompatibilists try to formulate the consequence argument in ways that can explain this difference. These formulations are quite technical, and it seems that there are difficulties with any version of the argument because compatibilists can reasonably refuse to accept some of the premises that are necessary for the argument’s conclusion (Kapitan 2002). This does not mean that the consequence argument is indefensible, but it does show that it is surprisingly difficult to elucidate its intuitive content in any exact way. In what follows, we shall discuss how philosophers try to understand the kind of control necessary for free will depending on what they think of the compatibility of physical determinism and the ability to do otherwise, as well as the relevance of the latter for the issue of free will and responsibility.

5 Libertarianism and Control

Let us assume that the consequence argument is valid, and hence free will is impossible in a deterministic universe. This means that if our universe is deterministic then free will is an illusion. Philosophers who believe that this is the case are called *hard determinists*. The main challenge for hard determinists is to explain what consequences of the alleged nonexistence of free will may have on moral responsibility, our social and legal practices, our evaluative thinking, and in general, our view of ourselves as agents (Pereboom 2001). However, since this chapter is about free will and not about what its denial might entail, we set these issues aside.

Philosophers who believe that our will can be free because our universe is not deterministic are called libertarians. They must answer at least two problems in order to make libertarianism a plausible philosophical view. First, they have to explain how free will is *possible* in nondeterministic universes. Second, they have to show how physical indeterminism is *relevant* to free will and responsibility. As we shall see, the second question is the harder one, but it is easier to classify libertarian views on the basis of their response to the first.

If determinism is false then the physical laws and the past states of the universe are logically com-possible with more than one future state of the universe. However, this is certainly not sufficient to grant that agents' will is free. As we have seen, libertarians think that the falsity of determinism is a condition of free will because it can only be up to me what I do if determinism is false. My heart beats at a certain pace in this moment. If determinism is false it is physically possible that it beats at a different pace even if the past states of the universe were the same. But this has little to do with my freedom of the will.

Thus, in order to explain the possibility of free will libertarians must explain how what we do can be *up to us*, if determinism is false. Agents cannot have free will unless they are able to control how they act. Otherwise their actions would be pure accidents. It seems, however, that control requires determination. But if nondetermined events occur as accidents, how is free will possible in a nondeterministic world? How can we reconcile the lack of determination with the presence of agential control? In this section, we shall consider briefly some libertarian attempts for such reconciliation.

According to some libertarians free will requires the introduction of a special kind of nondeterministic causation tailor-made for the purpose of explaining free agency (O'Connor 2000). Ever since Hume, causation is usually understood as a relation between events. Roughly, one event causes the other if the second follows the first in virtue of some nomic regularity. However, we often talk about *agents*, rather than events, as causes of actions. And normally it is *someone*—and not an event—whom we hold responsible for causing

something to happen. In general, when an agent acts she causes something to happen. Thus, it sounds plausible that we take agents rather than some events to be the causes of their behavior. Since agents are persisting substances and not events, a libertarian can argue that agents act of their own free will whenever they themselves, and not some earlier event, cause what they do.

Other libertarians object that we cannot make clear sense of the idea of agent causation. When agents act something happens; the result of the agent causing something must be an event. An event is a dated particular, but the agent itself is not. If there is no prior event causing the action, it is mysterious why the action took place at the time when it did (Ginet 1990). Although this objection is popular, it is far from being conclusive. The question is not whether there is always an event that is necessary for the occurrence of the other; rather, it is whether there is any event that could be sufficient. Perhaps it is true that some event must give occasion to the performance of an action. When it occurs to me that the traffic light has changed to red from green, this gives reason for me to push the break and hence explains the timing of my action. But without my contribution as a free agent, the push would not have occurred.

However, if we understand agent causation in this way, it is hard to see why only persons can agent-cause their actions. As far as ordinary language is concerned, it is as natural to say that "The bomb caused the damage" as it is that "The explosion caused the damage." Persisting substances can be causes by possessing certain powers and exercising them in specific circumstances. We may identify the cause as the substance or as the exertion of a power at a certain time in certain occurrent circumstances. In the former case, we identify the substance as the cause; in the second, we identify the occurrences, that is the events as causes. But even if it is natural to talk about persisting objects as causes they are not all responsible agents who have free will. An agent-causalist may want to respond that what distinguishes persons from other substances is that they agent-cause a special sort of mental event: their own choices. This sounds plausible, but then the relevant question concerns the nature of choice, and not agent-causation as a peculiar form of causal control (Lowe 2008).

An alternative "event causal libertarian" account of nondeterministic control relies on the notion of reasons. Most of the time when we explain agents' actions we do this with reference to their reasons: agents behave in the way they do because they act for this or that reason, and it is in this sense that their actions are not accidents. The crucial issue here is, of course, what it means to act for a reason. Some philosophers think that when we explain agents' actions in terms of reasons *for* which they act the explanation is teleological. However, it is more common to understand reasons in causal terms: reasons are agents' beliefs, intentions, desires that cause their behavior. It seems to follow that if

agents' reasons fail to causally determine their actions, then they lose rational control over what they do. On this ground, many compatibilists argued that rational control requires the truth of psychological determinism and hence libertarian free will is impossible (Hobart 1934, Ayer 1954/1982).

Some libertarians also claim that, given actual motives and reasons, agents cannot do otherwise (van Inwagen 1989). But they argue that agents can nevertheless be responsible because they have freely made themselves the kind of persons who have those particular motives and reasons. How is this possible if agents always act on their pre-existing reasons? It is possible because in some situations agents can have conflicting reasons for performing incompatible actions. If in these situations none of the reasons are "strong enough" to determine how to act, it is some genuinely indeterminate volitional effort which is causally responsible for the choice of the relevant action (Kane 1996). The interesting feature of such situations is that whatever agents do, they do it for a reason, and in this sense their action is not an accident.

However, this means that libertarians must restrict the class of genuinely free actions to those situations in which agents have nearly equally strong reasons for and against some action. Moreover, agents' decisions in such situations must have a character-forming effect that is sufficient to ground agents' responsibility for their later behavior. But this idea raises very difficult questions about self-determination and responsibility. If I have never been motivated to mistreat my children I can be responsible for this only if once I made some decision in dilemmatic situations that grounds my being the kind of person who does not want to mistreat them. However, I just cannot recall any such situation. Even if my choices in situations when I had conflicting reasons to act did contribute to the development of my character, it is only a very remote possibility that they have anything to do with the development of my parental preferences. Consequently, I cannot act in this respect of my own free will, and perhaps I'm not even "really" responsible for my parental behavior. For some, this seems to be a very high price for accepting a libertarian view of free will.

So we have seen that there have been three versions of libertarianism countenanced in recent debates. One is agent causation, another is a hybrid of agent and event causation, and the third is what one might call "Buridan's ass libertarianism," when the will can be free only in those situations in which reasons for and against an action are evenly balanced. All these face serious problems.

However, the major difficulty with all of these libertarian accounts of free will is to see exactly why physical—as opposed to social and psychological—determinism is incompatible with the kind of control that they believe necessary for free will and responsibility. Why would persons be unable to agent-cause their choices if physical determinism is true? And why cannot

they make “real choices” when they face conflicting reasons for actions or exercise volitional effort in a deterministic universe? Certainly, the relevant phenomenology of how I choose or make hard practical decisions will not be affected by a truth about physical determinism.

Moreover, the falsity of physical determinism itself cannot explain how I can agent-cause my choice or how, by exercising some volitional effort, I can choose for one reason rather than for another. If the nature of laws that govern the evolution of events at the sub-personal level are relevant for the exertion of such abilities, then in a universe in which basic physical laws are nondeterministic it is the objective chance of the exertion, rather than the exertion itself, that is determined “from below” so to say. Still, even in the nondeterministic case it is in virtue of the evolution of physical events that I choose and act as I do, and in this sense my actions are not “up to me.” If our mental powers and their exertion are the consequences of what happens at the sub-personal level then we cannot have free will or genuine agency. But if they are not, it is hard to see why the truth or falsity of determinism at the basic physical level is relevant for the possession and exertion of the abilities that constitute free will.

6 Semi-compatibilism and the Manipulation Argument

Many contemporary compatibilists are convinced that there is no conclusive refutation of the consequence argument. Rather, they think that the question of free will and responsibility should not be contingent upon how the debate about alternative possibilities turns out. As we have seen, freedom of the will is usually understood as the special form of control agents must be able to exercise over their own behavior in order to be responsible. Both incompatibilists and traditional compatibilists assume that such control is possible only if agents can do otherwise than they actually do. We may call this ability “regulative control.” Some compatibilists think that we need not have regulative control in order to be free and responsible. It is sufficient to have “guidance control” that requires only that agents somehow actively contribute to the production of their own behavior. They need not be able to do otherwise (Fischer 1994).

This view, according to which responsibility does not require alternative possibilities and it is for this reason that free will is compatible with determinism, is called semi-compatibilism. There are at least two important questions semi-compatibilist accounts of free will must answer. First, they must show that responsibility is indeed compatible with the absence of alternative possibilities. Second, they must identify the conditions that ground agents’ responsibility independently of their ability to do otherwise.

Semi-compatibilists argue for the irrelevance of alternative possibilities by introducing situations in which agents seem to be responsible even if they cannot do otherwise. Imagine a situation in which agent A makes a choice, but there is another agent B in the background who, had A been disinclined to choose as she does, would ensure that she choose so. Thus, the agent cannot avoid choosing to do what she does and hence she cannot act otherwise because her action is overdetermined in a special way. Actually, the action has been produced by the agent's own choice, but if she were about to choose otherwise, someone (or something) else would have ensured that she chooses to do what she actually does (Frankfurt 1988; Fischer 2002).

Semi-compatibilists claim that if the intervention does not occur because it is pre-empted by the agent's own choice then she is responsible for what she does. Although almost everyone agrees with this, it is contentious whether the agent indeed lacks alternative possibilities in the relevant situations. It is important to see that in the original example the overdetermination is only potential, and that the alternative causal process can become activated only as a response to the agent's inclination to choose otherwise. So first, it seems that even if the agent cannot choose to do otherwise some possibilities must be open to her before she makes the choice. As a response, it can be claimed that such possibilities, even if present, are irrelevant for responsibility because they do not enhance the agent's control. Another objection to the examples claims that the inactive presence of the alternative causal routes is not sufficient to deprive the agent of her ability to do otherwise. One can argue that the alternative causal route can ensure that the agent will make a particular choice only by depriving her of the ability to choose otherwise, which she retains in the absence of the intervention. And this is exactly that grounds her responsibility.

Let us assume that these objections can be answered, and agents indeed can be responsible even if they are unable to do otherwise. What makes then agents responsible? What are the conditions under which they are responsible for their actions? According to some semi-compatibilists, persons are responsible if they possess some generic moral capacities (Wallace 1998). According to others, agents are responsible if their actions are produced by some reasons responsive psychological mechanism (Fischer and Ravizza 1998).

However, there are difficulties with both proposals. On the one hand, I can have my generic rational and moral capacities while I'm asleep, but I do not omit actions of my own free will in that state. Thus, having some *generic* capacity is too weak to ground agents' responsibility for their particular behavior. On the other hand, we are often responsible for our irrational actions. But it is difficult to understand how they can be the results of the *specific operation* of some reasons responsive mechanism. Such problems with specifying the rational capacities necessary for responsibility might eventually prove that

the relevant notion of reasons responsiveness requires the ability to do otherwise. It is an old compatibilist idea after all that agents are responsible only if they would act differently if they had different reasons or motives. But this is an interpretation of alternative possibilities, and not a claim about their irrelevance (Huoranszki 2011).

Finally, if agents' reasons themselves are not up to them, perhaps we should not hold them responsible even if what they do is responsive to their reasons. As a response to this last problem, semi-compatibilists admit that reasons responsiveness is not sufficient for agents' free will and responsibility. Agents must also possess some reflective capacities that make the re-evaluation of their own motives possible. Further, they can be responsible only if some historical conditions are satisfied, which can ensure that their values and capacities have been formed in some "normal way" (Mele 1995; Fischer and Ravizza 1998). For instance, deprived childhood or subjection to severe indoctrination is not compatible with freedom and responsibility.

However, there are arguments that aim to show that the possession of the appropriate reflective capacities and the satisfaction of historical conditions cannot be sufficient for free will and responsibility. It is obvious that direct brain manipulation should not be allowed as a "normal way" to produce the reasons on which the agent acts. But some incompatibilists—the so-called source-incompatibilists—argue that physical determinism works in exactly the same way as manipulation does (Pereboom 2001). According to the so-called manipulation argument, from the point of view of agents' responsibility, there is no relevant difference between direct brain manipulation and the determination by the remote physical past.

The manipulation argument might be resisted, however. First, whether or not there is a relevant difference between direct brain manipulation cases and ordinary physical determinism is contentious at least for two reasons. Primarily, because the ordinary notion of manipulation presupposes an agent who intentionally misuses another, and hence what explains our intuition that the manipulated agent is not responsible is that the responsibility "passes over" to the manipulator. Thus, responsibility in such cases does not vanish, it is just "relocated." But the manipulation argument aims to prove that if physical determinism is true, there is no responsibility whatsoever. Moreover, direct brain-manipulation does not require the application of deterministic processes, thus it is unclear why determinism itself would be sufficient for undermining agents' free will and responsibility (Mele 2008). Second, even if we agree that physical determinism works like certain cases of manipulation do, we can deny that every form of manipulation is sufficient for exempting agents from responsibility. For instance, learning from others "manipulates" my reasons, but it is hard to see why this would affect my responsibility (Dennett 1984).

7 Traditional Compatibilism and the Conditional Analysis

Although many compatibilists believe that rejecting the alternative possibility condition is the most promising way to reconcile free will and determinism, it has remained questionable whether any example succeeds to demonstrate that agents can be responsible without the ability to do otherwise. More importantly, one can remain skeptical about whether reasons responsiveness itself, without the ability to act otherwise, can ever be sufficient for understanding free will. But this does not mean that we must reject compatibilism in general. Many traditional compatibilists agree with the incompatibilists that free will is impossible without alternative possibilities.

One such account is the so-called conditional analysis of free will. As we have seen, when agents act freely they must be able to control their behavior in some specific way. And the relevant kind of control may be best expressed with the help of a conditional: if the agent chose otherwise, she would also do otherwise. This conditional can be understood exactly as an interpretation of the ability to do otherwise. An agent has free will only if she can do otherwise, and she can do otherwise only if she has the power to do otherwise in the sense that she would do otherwise if she chose so. *Prima facie* at least, the conditional analysis of free will can capture well both the control condition and the alternative possibility condition of responsibility (Vihvelin 2004; Huoranszki 2011).

The conditional analysis is compatibilist as far as physical determinism is concerned. As we have seen, the idea behind the consequence argument is that if what we do now is the consequence of the past and the laws, and we are powerless with regard to the past and the laws, then we are also powerless to avoid our present actions. But this is simply false if we analyze powers in terms of conditionals. Obviously, from the fact that the remote past could not be made different by my choice *now* (and in this sense I'm powerless with respect to the past and the laws) it does not follow that my present action would not be different if I had chosen so. But one might ask do I have the power so to choose now? Given the logic of the concept of power, the answer must be that I do. Salt has the power to dissolve in water whether it is put in water or not, even if its not being put in water is determined by the past and the laws. Similarly, the agent has the power to do other things than what she did. Thus the conditional analysis is not only able to capture the responsibility relevant sense of control, it can also explain on what grounds compatibilists can reject the consequence argument.

Of course, the conditional analysis must also face objections. Some argue that the responsibility relevant sense of control does not require that we would do otherwise if we chose so. Consider the hit-and-run driver who is responsible for not helping her victim. It may not be true that she would have saved the victim of the accident if she had chosen so. Nonetheless, she is responsible

because she could have done otherwise. However, such counterexamples are not decisive against the conditional analysis. First, if it was *impossible* to save the victim then the hit-and-run driver may be responsible for some other omission that she would have done if she had chosen to. For instance, she is responsible for not even checking whether she could help. Second, if there was only *a chance* to save the victim, then it is true that she might not have saved him *even if*, she had chosen to. But the “even if” conditional expresses exactly that driver was able to save the victim in the sense that she would have had a chance to save him if she had chosen so. Thus, the relevant ability to do otherwise is still best captured by the conditional that one would do otherwise if one chose so.

Another objection to the conditional analysis is that the conditional can be satisfied even if the agent cannot make the relevant choice. This objection is related to a more general problem about how possibilities and conditionals are connected. It might be true of me that I would swim as fast as the current Olympic champion if I had the physical constitution he has. But it is not true of me that I *can* swim so fast, because I cannot have the physical constitution he has. Similarly, it might be true of me that I would answer an important phone call if I chose so even while I’m sound asleep. But I cannot answer it in the sense relevant for my responsibility. This certainly suggests that the conditional analysis of free will must be supplemented with further considerations about the psychological and epistemic conditions in which we can make choices (Huoranszki 2011).

However, as we have seen earlier when we discussed the libertarian views of free will, a theory about the relevant psychological and epistemic conditions need not include any view on the issue of physical determinism. There are many interesting philosophical questions about the phenomenology of conscious choice as well as how the experience of such choices is related to events occurring in the brain and to our physical behavior. These problems are not less pressing than the traditional question of compatibility and control, but they require separate discussion. They are problems about how conscious processes can exist and be causally relevant in a physical world (Swinburne 2011).

To sum, I have tried to show the following: (a) The best attempts to develop a libertarian account of free will have run into the ground not so much because they cannot provide an adequate account of the abilities necessary for free will, but because they failed to show why the falsity of determinism at the fundamental physical level is relevant for the possession of such abilities. (b) Determinism on the human level—determinism in social, psychological, and at least some biological terms—is very probably incompatible with the ability to do otherwise. (c) Despite (b), it is not clear that determinism at the basic physical, and hence the sub-personal level is incompatible with the ability to do otherwise. (d) This is so because a “powers” sense of “could have done otherwise” is compatible with sub-personal determinism.

Bibliography

- Ayer, A. J., 1954. "Freedom and Necessity." In G. Watson, ed., 1982, pp. 15–23.
- Cahn, S. M., 1967. *Fate, Time and Logic*. New Haven: Yale University Press.
- Clarke, R., 2003. *Libertarian Accounts of Free Will*. Oxford: OUP.
- Craig, W. L., 1987. *The Only Wise God. The Compatibility of Divine Foreknowledge and Human Freedom*. Grand Rapids: Baker Book House.
- Dennett, D. C., 1984. *Elbow Room: The Variety of Free Will Worth Wanting*. Cambridge, MA: MIT Press.
- Fischer, J. M., 2002. "Frankfurt-Type Cases and Semi-Compatibilism." In R. Kane, ed., *The Oxford Handbook of Free Will*. Oxford: OUP, pp. 281–308.
- , 1994. *The Metaphysics of Free Will*. Oxford: Blackwell.
- Fischer, J. M. and Ravizza, M., 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: CUP.
- Frankfurt, H., 1988. *The Importance of What We Care About*. New York: CUP.
- Ginet, C., 1990. *On Action*. Cambridge: CUP.
- Hobart, R. E., 1934. "Free Will Involving Determination and Inconceivable without It." *Mind*, 63, pp. 1–27.
- Huoranszki, F. 2011. *Freedom of the Will. A Conditional Analysis*, New York: Routledge.
- , 2002. "Fate, Freedom and Contingency." *Acta Analytica*, 17, pp. 79–102.
- Kane, R., 1996. *The Significance of Free Will*. New York: OUP.
- Kane R., ed., 2002. *The Oxford Handbook of Free Will*. Oxford: OUP.
- Kapitan, T., 2002. "A Master Argument for Incompatibilism?" In R. Kane, ed., 2002, pp. 127–57.
- Lowe, E. J., 2008. *Personal Agency: The Metaphysics of Mind and Action*. Oxford: OUP.
- Mele, A. R., 2008. "Manipulation, Compatibilism, and Moral Responsibility." *Journal of Ethics*, 12, pp. 263–86.
- , 1995. *Autonomous Agents. From Self-Control to Autonomy*. Oxford: OUP.
- O'Connor, T., 2000. *Persons and Causes: The Metaphysics of Free Will*. New York: OUP.
- Pereboom, D., 2001. *Living without Free Will*. Cambridge: CUP.
- Strawson, G., 1994. "The Impossibility of Moral Responsibility." *Philosophical Studies*, 75, pp. 5–24.
- Swinburne, R., ed., 2011. *Free Will and Modern Science*. Oxford: OUP.
- van Inwagen, P., 1989. "When Is the Will Free?" *Philosophical Perspectives*, 3, pp. 399–422.
- , 1983. *An Essay on Free Will*. Oxford: Clarendon.
- Vihvelin, K., 2004. "Free Will Demystified: A Dispositionalist Account." *Philosophical Topics*, 32, pp. 427–50.
- Wallace, R. J., 1998. *Responsibility and the Moral Sentiment*. Cambridge, MA: Harvard University Press.
- Watson, G., ed., 1982. *Free Will*. Oxford: OUP.

26 Knowledge

Bryan Frances and Allan Hazlett

This chapter recommends four topics in the theory of knowledge: skepticism, knowledge attributions, the nature of knowledge, and the value of knowledge. These topics (among others) have taken center stage in contemporary analytic epistemology.

1 Skepticism

You can't get far on a theory of truth unless you tackle the liar paradox. The reason is that that paradox seriously threatens the truth of some of the most fundamental assumptions one is naturally inclined to make regarding truth and meaning. For analogous reasons, one can't get far in epistemology unless one deals with the skeptical paradox.

Skeptical arguments come in various forms, but in recent philosophy almost all of them have this one:

- (a) In order to know *P*, you have to meet condition *C*.
- (b) You don't meet condition *C*.
- (c) So, you don't know *P*.

Here are just two candidates for *C*:

C1: You know, or are in a position to know, or justifiably believe, that it's not the case that almost all your ordinary beliefs aren't true because you're a brain in a vat being fed electrical signals so as to generate the mere illusion of having a normal body and living a normal life.

C2: You know, or are disposed to know, or justifiably believe, that it's not the case that almost all your ordinary beliefs aren't true because there are no trees, shoes, or other ordinary objects, as the paradoxes about material objects prove (there are no shoes, although there are "particles arranged shoe-wise," and your beliefs are about the former instead of the latter).

P is chosen to be a substantive, contingent, empirical claim almost all of us think is pretty well known, such as “I own three pairs of shoes.” The reason these arguments merit close attention—and are considered paradoxical—is that, with P and C specified in this way:

- There are good reasons for (a).
- There are good reasons for (b).
- There are good reasons against (c).
- There are good reasons for the validity of the argument (a)–(c), so there are good reasons to think that it’s impossible for (a) and (b) to be true while (c) is false.

We tend to think that perception and elementary reasoning are pretty good at generating evidence or justification fully adequate for knowledge (when combined with true belief). We do seem to know at least *some* substantive, contingent, empirical claims, by using perception and ordinary reasoning, like “I own three pairs of shoes.” So, we are highly confident in the falsehood of (c).

The reason (a) is so plausible is that it is so easy to *reliably reason* from “I own three pairs of shoes” to “It is not the case that C holds,” for either C1 or C2. After all, if you know (or justifiably believe) that you own three pairs of *shoes*, then it’s the easiest thing in the world to see that it simply has to be the case that *there are shoes*, which is obviously inconsistent with both C1 and C2. Often this is put in terms of the *Closure Principle*: if you know P (e.g. that you own three pairs of shoes) and you know that P entails Q (e.g. that your owning three pairs of shoes entails that there are shoes), then you know (or are in a position to know) Q (e.g. that there are shoes). If we can reliably reason from “There are black holes in the Milky Way galaxy” to “There are black holes,” then surely we can reliably reason from “I own three pairs of shoes” to “There are shoes.” Reliably reasoning from “There are shoes” to the falsity of C1 and C2 looks just as elementary.

The reason (b) is so plausible is that it seems like we don’t satisfy C1 or C2. Take C1 first. What could you possibly *do* to show that you’re not a brain in a vat (BIV)—with or without three pairs of shoes? Well, someone might say, mere brains don’t have feet, and you can clearly *see* your feet. But that so-called seeing might be an illusion manufactured by the mad scientists who are controlling your brain. You might say that it’s impossible for scientists or anyone else to create this illusion. But that “fact” about science might be an illusory bit of “information” fed to you by just those scientists. It doesn’t take a great deal of reflection before you see the plausibility of the idea that you can’t do anything to show that you’re not a BIV—at least not in any ordinary way of doing something to prove a point. In addition, in all probability there is little you can do to show that the paradoxes about material composition fail

to show that there are no shoes in our universe. However, C1 and C2 seem a bit different on this score. Whereas it seems that it was *impossible* for you to do anything to rule out the BIV possibility, we have the feeling that it's possible for you to do something to rule out the no-shoes possibility indicated in C2, as you could "just" give a knockdown argument that refutes the philosophers and physicists who say that there are no composite objects, including shoes. But if you don't have the ability to do anything to rule out those metaphysical theories, then it looks like you can't do anything to rule out the no-shoes possibility mentioned in C2.

So is there some other way you can know that you're not a BIV and that the paradoxes of material composition don't show that there are no shoes? You might appeal to the commonsensical point that knowing something and proving it are two different things—that someone can know something even if she cannot prove it. Sometimes we have the ability to tell that something is the case, but not the ability to prove that it's the case—think of how you have the ability to recognize people by their faces, but would be hard-pressed to prove that your identifications were correct. However, this won't help with C1 and C2. Perhaps an elite professional metaphysician has the resources to tackle the paradoxes of material composition, and so has the ability to tell whether they do or do not show that there are no shoes, but most of us don't have that ability. And the situation is even worse when it comes to C1, for it doesn't seem possible that *anybody* could have the ability to tell whether she was a brain in a vat. So (b) certainly seems true.

The argument (a)–(c) certainly appears to fit *modus tollens* perfectly, and since that's a deductively valid form of reasoning, it seems impossible for (a) and (b) to be true while (c) is false.

Hence, we have a paradox: a collection of four claims that are individually plausible yet collectively inconsistent. Here are the four claims:

- (a) is true
- (b) is true
- (c) is false
- It's impossible for (a) and (b) to be true while (c) is false.

So there must have been some mistake in our reasoning. The reasoning offered above in favor of those four claims is profoundly misleading in some way that is hard to detect. In fact, it is so hard to detect that despite centuries of dedicated investigation epistemologists have come to no agreement regarding which of the four claims is false. Our task is to discover which claim is false, and to explain why.

Since there are four claims in our inconsistent set, the four ways of avoiding the inconsistency involve denying one of the four individual claims. That

is, we have *just four* options in responding to the paradox: hold that (a) is false, hold that (b) is false, hold that (c) is true, or hold that the argument isn't valid. One can hardly claim to have a thorough response to the paradox if one doesn't indicate which option one is adopting. *Skepticism* is the view that (c) is true. Let's consider the other possible thorough responses.

The *Neo-Moorean* says that (b) is false. The position gets inspiration from G. E. Moore, who would have said that (b) is false. On this view, we do know that we're not BIVs. Different philosophers give different reasons for rejecting (b). Moore (1939) would argue that we can know we're not BIVs by deducing that conclusion from some ordinary item of knowledge, such as "I own three pairs of shoes"; versions of this idea have been defended by Ernest Sosa (1991, 1999a, 1999b) and James Pryor (2000). For *reliabilist* Neo-Mooreans (like Sosa), the mere fact that my belief-forming faculties are reliable—the mere fact that they usually yield mostly true beliefs—is enough to ensure that most of my beliefs amount to knowledge, including my belief (should I come to form it) that I'm not a BIV. On this view, we do not need to do anything to show that our faculties are reliable, to ensure that most of our beliefs amount to knowledge; for reliabilists it is sufficient that our faculties *are* reliable. For criticism of this approach, see Stroud (1994) and Fumerton (1995). Alternatively, Jonathan Vogel (1990) argues that we can know we aren't BIVs via an inference to the best explanation; Berent Enç (1990) offers criticism of that view.

The *closure denier* says that (a) is false. For example, Robert Nozick (1981) argues that we can know that we have shoes even though we don't and can't know that we aren't shoeless BIVs. In so doing, these philosophers reject the Closure Principle (mentioned above): if you know P and know that P entails Q, then you know (or are in a position to know) Q. On the one hand, this is an attractive position: we get to say, with common sense, that *of course* we know that we own three pairs of shoes, but we also acknowledge the skeptic's achievement by admitting that we can't know that we aren't shoeless BIVs. But then again, isn't it incredibly easy to deduce, from your knowledge that you have three pairs of shoes that fit your feet quite well, that you have shoes, and so come to know that have you shoes? What inference could be easier? And isn't it equally easy to then deduce that you're not a shoeless BIV, which by definition has no shoes? Such considerations make the rejection of (a) and the Closure Principle look awkward at best. The places to start studying: Dretske 1970, Nozick 1981, Kvanvig 2006, and Kripke 2011.

The last way of dealing with the skeptical paradox, rejecting the validity of the (a)–(c) argument, introduces the topic of the next section: epistemological attributions. By studying how we *talk about* knowledge we may not only get a new way to respond to the skeptical argument but a better understanding of what knowledge really is.

But before we get to that issue it is worth mentioning that epistemologists have recently been studying another kind of skeptical problem. The argument goes something like this: you start out believing *P*, you then acquire some information that gives you powerful reason to think that your belief is false or your basis for your belief is seriously flawed; you don't have any evidence that can "counter" that worrisome information; so, you don't know *P* (even if you started out knowing *P*). For instance, *P* might be "We have free will," "God doesn't exist," or "Truth is a consistent notion," but you know that there are loads of people smarter and much better informed than you are who think your belief is false. Alternatively, the worrisome information you acquire might be the claim that most people's opinions on *P* are typically greatly exaggerated, and there is no reason to think you are an exception. Or *P* might be some political opinion of yours, and the problem is that people almost always reach their political opinions through nonevidential means, and there is no reason to think you are an exception. On this kind of skeptical problem, see Feldman 2005, Frances 2010 and forthcoming, Kornblith 2010, and White 2010.

2 Knowledge Attributions

A story condensed from DeRose (2009) is an excellent starting point into the literature on knowledge attributions:

Thelma, Louise, and Lena are friends who all work in the same office. Today is their day off, but, before getting an early dinner together, they decide to walk up to the office to pick up their paychecks. Thelma and Lena are also interested in finding out whether their colleague John is at work, as they are involved in a small office bet with a couple of other workers over whether John would show up today. As they pass the door to John's personal office, they see his hat hanging on the hook in hallway, which, in their long experience, has been a sure-fire sign that John is in fact at work. Satisfied that John is at work and that Thelma and Lena, who bet that he would be, are in a position to collect their winnings from some other office workers, the three friends pick up their checks, go out to dinner together, and then part company, Thelma going to a local tavern to meet other friends, and Louise and Lena each heading in different directions to go home.

Thelma meets a friend at the tavern and says to him, "Hey, John was at work today. Pay up!" When her friend asks, "How do you know?" Thelma describes the above facts. Then Thelma says, "Lena knows that John was there, too, as she was with me."

At the same time Louise is stopped by the police on her way home. They are conducting an extremely important investigation and are seeking to determine whether John was at work that day. When the police ask her whether she could testify that John was at work, Louise replies, "Well, no, I never saw him. I could testify that I saw his hat hanging in the hall, which is a very reliable sign that he's at work. But I suppose John could have left his hat on the hook when he went home some previous day." When the police ask Louise whether Lena might know whether John was in, Louise replies, "No. She has the same reasons I have for thinking John was there, but, like me, Lena doesn't know that John was there."

It seems that when Thelma says at the tavern "Lena knows that John was there," her claim is true. But it also seems that when Louise tells the police "Lena doesn't know that John was there," her claim is true. But that looks impossible, as they seem to be asserting contradictory claims. Theorists of knowledge attributions have the task of explaining why both Thelma's "Lena knows that John was at work" and Louise's "Lena doesn't know that John was at work" are both conversationally appropriate even though they certainly appear to be inconsistent with one another.

Contextualists adopt the most straightforward explanation: the reason the two assertions are conversationally appropriate is that they are both *true* (Cohen 1988, 1999; DeRose 1995, 2009; Lewis 1996). And that means that for at least some sentences of the form "S knows P," there are two tokens of the sentence that concern the same S and P (and time) where one token is true and another is false. This can happen, according to contextualism, when the two tokens are uttered in different contexts. So "S knows P" is akin to "S is tall," which can have two tokens with different truth-values depending on the standards that are operative in two different contexts (e.g. a single person can, at one time, be tall for a high school student but not tall for a basketball player).

Contextualists need an explanation of *how* tokens of "S knows P" get different truth-values, and this explanation will have to pay careful attention to semantic issues; this is one reason why the philosophy of language has become important to epistemology. One idea (mentioned above) is that in different conversational contexts there are different epistemic *standards* a belief has to meet in order to be called "knowledge": in Louise's police context the standard is high, in Thelma's tavern context it is low, and Lena's belief meets the low but not the high standard. Again, compare this to how "Chris is tall" can be true relative to one context (in which we are discussing high school students in general) but false relative to another (in which we are discussing only the students on the high school basketball team). But there are alternative ways of understanding the context-sensitivity of "knows," including by

appeal to salient alternatives (Lewis 1996) and salient contrast (Schaffer 2005, 2007).

According to the contextualist, whether or not *my* use of “S knows P” is true will depend not only on S (e.g. whether she believes P, what evidence she has, whether her belief was formed reliably) and on P (whether P is true), but will depend on the context surrounding *me*. It isn’t hard to see how these ideas might apply to the skeptical paradox. When we are strongly inclined to reject (c)—that is, to reject the sentence “So, you don’t know P”—contextualists say we are in a low-standards, “everyday” context in which (c) is false (or that we are imagining ourselves in such a context). But when we are attracted to the skeptic’s premise (b) —“You don’t meet condition C”—contextualists say we are in a high-standards, “skeptical” context in which (b) is true (or we are imagining ourselves in such a context). (Alternatively, (b) might be without truth-value in that context because of some kind of indeterminacy; see DeRose 2004). In the low-standards context I speak the truth when I utter the sentence “I know I have three pairs of shoes,” even though this sentence would be false (or indeterminate) were I to utter it in a high-standards context. The skeptical conclusion, in a high-standards context, is compatible with the truth, in a low-standards context, of our ordinary claims to knowledge. For criticism of this response to the skeptical paradox, see Sosa (2000).

Contextualism is a theory about the meaning and use of “knows,” and it has several rivals in the theory of knowledge attributions. According to *subject-sensitive invariantism* (Hawthorne 2004; Stanley 2005; Fantl and McGrath 2009), whether a use of “S knows P” is true depends not only on such “epistemic” factors as S’s evidence or the reliability of S’s methods of belief-formation but also on S’s *practical circumstances*. (Subject-sensitive invariantists thus endorse “pragmatic encroachment,” which will be discussed in the next section.) Contextualism says that the details of the *attributor* of “S knows P” make a difference as to its truth conditions; subject sensitive invariantism says that the practical circumstances of the *subject* S make a difference as to whether S knows. However, a contextualist could adopt (something like) both theses by saying that the truth-value of “S knows P” depends on both subject and attributor factors (Lewis 1996; DeRose 2009). Alternatively, *relativism* says that the truth-value of “S knows P” depends on the person *evaluating* the sentence (MacFarlane 2005). Finally, *Griceans* (Rysiew 2001; Hazlett 2009) deny that the truth condition for “S knows P” vary with context, and account for the difference in conversational appropriateness by appeal to Grice’s theory of conversational implicature. For instance, perhaps when Louise says to the police that Lena doesn’t know that John was at work she speaks falsely—Lena really does know—but correctly conveys the information that the police will not be able to rely on Lena’s testimony in court. Louise is trying to be informative for the police: although they asked whether Lena knows, what they are really after

is whether Lena *saw John at work*, and were Louise to say that Lena knows, it would imply something false: that Lena saw John at work.

The task for the theory of knowledge attributions is to offer a detailed response to the various cases in a convincing way. In particular, one needs a well developed response to as many of the cases as possible (for a good portion of them, see Hawthorne 2004; Stanley 2005; and Rysiew 2011).

3 The Nature of Knowledge

The skeptical problem concerns the *scope* of our knowledge: it concerns *what* we can know, or *how much* we can know. A distinct, though related, problem concerns the *nature* of knowledge; this problem concerns what knowledge *is*. (Note that this is also distinct from, though related to, a question discussed above, about the meaning and use of the word “knows.”) Analytic philosophers have explored the nature of knowledge primarily by attempting to articulate necessary and sufficient conditions for knowledge. A minority look elsewhere, to the genealogy of the concept of knowledge (Craig 1990; Kusch 2009; Dogramaci forthcoming), for illumination.

According to the *tripartite analysis of knowledge*, first articulated by Socrates in the *Meno*,

S knows that p iff (i) p is true, (ii) S believes that p, and (iii) S is justified in believing p.

We'll comment on each of the three conditions. Knowledge, according to condition (i), is *factive*: being in such a factive state with respect of some proposition P entails that P is true. (Note well that the view that knowledge is factive does not entail that the truth is transparent to us, for it is compatible with this view that we rarely, if ever, know whether we know.) Condition (i) is sometimes defended by appeal to a linguistic thesis about “knows” (see Hazlett 2010). We should keep in mind that cultural or historical relativists about truth can embrace the necessity of condition (i): that people once knew that the earth was flat is no counterexample to the necessity of the truth condition, if at the time they knew that proposition, it was true.

According to condition (ii), knowledge is a species of belief. This view can be challenged by appeal to the following kind of case (Radford 1966; Williams 1973): an intellectually cautious student correctly answers a set of exam questions, while suspending judgment on the contents of her answers; the student knows the answers to the exam questions, but she does not believe those answers. This suggests an alternative conception of knowledge, on which to know is to possess (without necessarily believing) true information (Williams

1973), or perhaps to have the ability to access true information in action (Ryle 1949). The issues at play here are as much issues about the nature of belief as about the nature of knowledge; compare a conception of belief on which it is that propositional attitude that combines with desire to produce action, on which the student in the above example does believe her answers, with a conception of belief on which it entails conscious, affirmative judgment, and on which the student does not believe her answers. It is agreed that knowledge requires the possession, in some sense, of true information. The tripartite analysis understands "possession" as belief; others (Radford and Williams) think this requires too much; still others think this requires too little, since knowledge requires certainty (Ayer 1956; Unger 1975).

Condition (iii) is controversial in two ways. First, some argue that true belief is sufficient for knowledge (Sartwell 1991; see Hazlett 2010). A jealous husband knows about his wife's affair when he truly believes that she is having an affair, regardless of whether his belief is justified (Williams 1973). Alternatively, you might think that "knows" is polysemous such that on one sense of "knows" it is synonymous with "true belief." Second, some theories of knowledge, while not endorsing the thesis that true belief is sufficient for knowledge, do not require justification for knowledge (Goldman 1967; Nozick 1981).

The fate of condition (iii) depends on the nature of justification. Some suggest a *dialectical conception of justification*, on which being justified in believing that P is a matter of being able to give reasons for P, or of being able to justify one's belief that P to others, or of being able to avoid criticism for believing P (Ayer 1956; Wittgenstein 1969). Two sorts of considerations speak against appealing to this conception of justification in our analysis of knowledge. First, we may have occasion to attribute knowledge to creatures who fall outside the dialectical "space of reasons," for example young children and other non-linguistic animals. Second, we may find ourselves unable to justify our ordinary beliefs against the skeptic's challenge, thus anti-skeptical epistemology will require a non-dialectical conception of justification (Sosa 1991).

We might then opt for an *evidentialist conception of justification*, on which being justified in believing that P is a matter of having sufficient evidence for P. According to a prominent evidentialist account of knowledge, knowledge is true belief that is well-founded, where a belief in P is *well-founded* iff it is properly based on evidence that justifies P (Conee and Feldman 2004).

Reliabilists (see above) reject this account of knowledge, either in favor of the view that justification is not necessary for knowledge (see above) or in favor of a *reliabilist conception of justification* (Goldman 1979; Plantinga 1993; Sosa 1991), on which being justified in believing P is a matter of one's belief in P being formed (and sustained) in a reliable way. This leaves open the nature of reliability; one possibility is a modal account (Nozick 1981; Sosa 2001) that defines reliability in terms of sensitivity (S's belief that p is sensitive iff were it

the case that $\sim P$, then S wouldn't believe P) or safety (S's belief that P is safe iff S would not easily believe P when P is false).

At the heart of the dispute between evidentialists and reliabilists is a disagreement about whether it is possible for someone to know P without having reasons for believing P. There is an important sense of "having reasons" on which reliabilists deny that knowing requires having reasons. Epistemologists have discussed cases in which someone's belief in p is reliably formed and sustained, but where (in the aforementioned sense) she does not have reasons for believing P, for example a savant endowed with a highly reliable faculty of mathematical intuition, whose mathematical beliefs are by hypothesis not based on evidence. The beliefs of the savant are reliably formed and sustained, no matter how you understand these notions, but not well-founded. Shall we say that she knows? If yes, then our sympathies lie with some form of reliabilism. If no, then with some view more akin to evidentialism.

The tripartite analysis of knowledge is rejected by most contemporary epistemologists, on the basis of the following kind of counterexample (Gettier 1963):

Smith and Jones are the only applicants for a certain job. Smith has strong evidence (testimony from the president of the company they have applied to) that Jones will get the job, and also (perceptual evidence) that Jones has ten coins in his pocket. From these two propositions Smith deduces that the man who will get the job has ten coins in his pocket. However, it turns out both that Smith will get the job (the president was mistaken) and that Smith happens to have ten coins in his pocket (although he doesn't know this).

Smith's belief (that the man who will get the job has ten coins in his pocket) is true and justified; however, intuitively, Smith does not know that the man who will get the job has ten coins in his pocket. The tripartite analysis is therefore false. Coming up with a satisfactory account of the nature of knowledge that is not subject to this kind of counterexample is known as the *Gettier problem*.

Why does Smith fail to know? You might think that he fails to know because his belief is deduced from a falsehood; this thought forms the basis of the "no false lemmas" approach to the Gettier problem, on which S knows that P only if S's belief that P is not deduced from any falsehoods, but many have thought that this diagnosis fails to capture the essence of Gettier's case. Consider another case (Goldman 1976):

Henry is unknowingly driving through "fake barn country," where most of structures in the fields are barn facades, which from a distance look like real barns; the rest of the apparent barns are real. Driving past one of the

rare real barns, Henry looks out the window, sees the barn in the distance, and forms the belief that there is a barn in the field.

Henry's belief (that there is a barn in the field) is true and justified; however, intuitively, Henry does not know that there is a barn in the field. The tripartite analysis is therefore false, and in this case Henry does not deduce his belief from any falsehood: his belief is simply based on (or is a constituent of) his perception of the barn in the field.

The case of Henry suggests that whether someone knows depends crucially on her *environment*. Had Henry formed the belief that there's a barn in the field while driving through a more normal environment, he would have known that there was a barn in the field. And both the case of Smith and the case of Henry suggest that knowledge is incompatible with certain kinds of *luck* (Nagel 1979; Pritchard 2005). The problem with the tripartite analysis is that it fails to take any of these facts into account.

Reliabilist virtue epistemologists (Sosa 2004; Greco 2010) have articulated a solution to the Gettier problem based on the idea that knowledge is a cognitive achievement. In general, we can distinguish between (mere) success and achievement: an archer is successful when her arrow hits her target, but her success is an achievement when her arrow hits her target as a result of her archery skills, if her successful shot manifests her competence at archery. We are asked to think of cognition as analogous to such an activity, and as admitting of the same distinction: my (mere) true belief is a success, but my knowledge is an achievement—a success that manifests my competence at forming true beliefs. Given this conception of knowledge, we can see why Henry's true belief falls short of knowledge: it is not an achievement of Henry's. (The same, *mutatis mutandis*, when it comes to Smith.) His belief is analogous to a competent archer's shot which reaches its target by luck, for example by being blown off course, and then back on course, by two chance gusts of wind. Such lucky success is to be distinguished from genuine achievement, where success is *due to* or a *manifestation of* the abilities of the successful agent.

Timothy Williamson (2000) proposes a more radical approach to the Gettier problem: a rejection of the project of analyzing knowledge in terms of belief and evidence, in favor of "*knowledge first*" *epistemology*, on which notions like belief and evidence are to be analyzed in terms of knowledge.

The Gettier problem suggests that whether someone knows depends crucially on her environment; other cases suggest that whether someone knows depends crucially on her *practical circumstances*, including the decisions she has to make and her preferences and concerns:

Train 1: John is waiting on a train platform as a train pulls up and stops. He has taken this particular train about twice a month for a couple years.

Someone standing next to him asks him “Do you know if this train stops at Melrose?” John has never taken the train to Melrose but he is certain he has heard the train personnel say “Melrose is the next stop” a great many times while riding this train; his memory is ordinary. So, he says “Yes, it stops at Melrose.”

Train 2: Everything is the same as in case 1 except that the people asking John are police officers with “Bomb Squad” on their uniforms who have run up on the waiting platform and announced that it’s imperative that they get to the Melrose station as soon as possible. They turn to John and ask him “Do you know if this train stop at Melrose?” John understands the stakes at hand. He says “I am pretty sure it does but you’d better ask the conductor.”

A natural explanation of why it’s fine, in Train 1, for John to say that the train stops at Melrose is that he knows that the train stops at Melrose, but an equally natural explanation of why it would be wrong, in Train 2, for John to say that the train stops at Melrose, is that he *doesn’t know* that the train stops at Melrose. If he *knew*, in Train 2, why wouldn’t he be able to appropriately tell the police that the train stops there?

According to the thesis of *pragmatic encroachment* (Hawthorne 2004; Stanley 2005; Fantl and McGrath 2009), whether someone knows that P depends on her practical circumstances: in Train 2, more evidence is needed for John to know that the train stops at Melrose, given that something serious seems to ride on the reliability of his testimony. Defenders of pragmatic encroachment have understood their view as a radical departure from epistemological orthodoxy, although this picture has been challenged (Grimm 2011). Alternatively, you might favor a Gricean account of the train cases (Hazlett 2009), on which John needs more than knowledge to answer the police affirmatively in Train 2.

Inquiry into the nature of knowledge has received criticism recently on the ground that it is illegitimately based on appeals to philosophical intuition, and in particular that it is illegitimately based on appeals to philosophical intuitions about the extension of the concept of knowledge. “Experimental philosophers” have argued that people’s intuitions about Gettier cases vary with presumably irrelevant factors such as ethnic background and gender (Weinberg et al. 2001; Buckwalter and Stich 2011), but these experimental results could not be replicated (Nagel forthcoming). Alternatively, the legitimacy of appeals to intuition can be problematized a priori by asking whether knowledge is a natural kind (Kornblith 2006). For if knowledge is not a natural kind, we might then question the philosophical, as opposed to anthropological or sociological, interest in limning the extension of some socially constructed concept of ordinary language.

An alternative to the method based on appeals to philosophical intuitions about the extension of the concept of knowledge is one based on appeals to evaluative premises, where the value of knowledge is a constraint on theories of the nature of knowledge (Kvanvig 2003). This project promises more than a limning of the extension of a socially constructed concept of ordinary language—it promises an account, at least, of something that we care about (in as much as there is a connection between what we care about and what is valuable). And when we return to the “traditional” methods employed in the theory of knowledge, we find the evaluative method has been with us all along: the tripartite analysis is articulated by Socrates in connection with the question of how knowledge is more valuable than (mere) true belief (see below), and regardless of whether the cognitive achievement conception of knowledge is well-motivated by intuitions about Gettier cases, which are the sort of intuitions to which we have just considered challenges, it seems well motivated by the idea that knowledge is more valuable than (mere) true belief (again, see below).

4 The Value of Knowledge

Many epistemologists maintain that knowledge is valuable. This claim, however, admits of numerous disambiguations.

First, we can distinguish between the claim that knowledge is *always* valuable and the claim that knowledge is *sometimes* valuable. Second, consider the claim that knowledge per se is valuable, that is that knowledge is valuable “as such”: being F is per se valuable when anything that is F is valuable just in virtue of being F. (Therefore, that knowledge per se is valuable entails that knowledge is always valuable.) Third, consider the claim that knowledge is *prima facie* valuable: X is *prima facie* valuable when X is valuable “at first glance” or “on its face.” Fourth, consider the claim that knowledge is *pro tanto* valuable: X is *pro tanto* valuable when X has a value that can be trumped by other values. Fourth, consider the claim that knowledge is *intrinsically* valuable: X is intrinsically valuable iff the fact that X is valuable supervenes on the intrinsic properties of X. (Intrinsic value contrasts with *extrinsic* value.) Fifth, consider the claim that knowledge is *finally* valuable: X is finally valuable iff X is valuable for its own sake. Final value contrasts with *instrumental* value: X is instrumentally valuable iff X is valuable for the sake of something else (wholly distinct from X). (Intrinsic value is therefore a species of final value, and instrumental value a species of extrinsic value.) Sixth, you might think that knowledge has (what we can call) *eudaimonic* value: X is eudaimonically valuable when X is valuable vis-à-vis wellbeing. So, for example, you might think that knowledge is *good* for the knower; if you think this, then you think

that knowledge has eudaimonic value. Seventh, you might argue that knowledge has *constitutive* value: X has constitutive value iff X is valuable in virtue of being part of a valuable whole, such as a good human life. (Constitutive value could naturally be understood as a species of final eudaimonic value, to be contrasted with instrumental eudaimonic value, which could be called “practical” or “prudential” value.) Eighth, you might want to say that knowledge has *epistemic* value, although the notion of “epistemic value” is obscure, as is its relationship to eudaimonic and constitutive value.

In contemporary epistemology, the most frequently discussed question concerning the value of knowledge is the *Meno question*: why, and in what way, is knowledge more valuable than true belief (if it is)? In the *Meno*, Socrates presents an answer based on the idea that knowledge is more secure than (mere) true belief (97e–98a). On its face, this suggests that the instrumental eudaimonic value of knowledge is always greater than the instrumental eudaimonic value of (mere) true belief, in virtue of the fact that knowledge is resistant to revision in a way that (mere) true belief isn’t. But this seems wrong. Imagine that a credulous nitwit is absolutely convinced of some proposition, on the basis of the testimony of his obviously unreliable guru. The proposition, as it turns out, is true. Suppose that a cautious thinker is reasonably confident of the truth of the same proposition, on the basis of careful consideration of the evidence that bears on the proposition. The (mere) true belief of the credulous nitwit might be much more secure, much more resistant to revision, than the knowledge of the cautious thinker: nothing will make the nitwit give up her conviction, while the cautious thinker will be moved to revise if evidence against the proposition emerges.

Contemporary epistemologists (Kvanvig 2003; Pritchard 2010) have articulated the *Meno* question in terms of epistemic value: why, and in what way, is knowledge more *epistemically* valuable than (mere) true belief (if it is)? As Duncan Pritchard (2010) argues, the value problem arises when we assume:

Truth value monism: True belief (and avoiding false belief), and only true belief (and avoiding false belief), has fundamental epistemic value.

Truth value monism does not entail that only true belief has epistemic value. It allows for the possibility of other epistemic goods, but only when the epistemic value of said goods can be explained in terms of the epistemic value of true belief. Thus the truth value monist might explain the epistemic goodness of the virtue of intellectual humility by appeal to the fact that (a) the intellectual humble person will be a more reliable believer, and (b) a more reliable believer is one more likely to acquire true beliefs. But truth value monism seems to imply that any true belief, whether amounting to knowledge or not, will be doing maximally well in terms of the fundamental epistemic good. Compare

two beliefs of mine, one an instance of knowledge, the other not an instance of knowledge, but merely a true belief. Both are equally good in terms of the only fundamental epistemic good, according to truth value monism. My knowledge differs from my (mere) true belief, though: perhaps the former is based on good evidence, while the latter is wishful thinking. But if truth value monism is right, the epistemic value of basing your beliefs on good evidence must be non-fundamental; we must say that basing your beliefs on good evidence is epistemically valuable because this makes it more likely that your beliefs will be true (or something like that). But now we cannot make sense of how my knowledge is more valuable than my (mere) true belief, in virtue of the former being based on good evidence: for my (mere) true belief already enjoys the epistemic good that basing one's beliefs on good evidence is said to bring. Non-fundamental epistemic value is "swamped" by fundamental epistemic value: when the latter is present, the former loses its appeal.

Compare a gold miner, who (like the Bond villain) loves only gold (see Zagzebski 1996). Gold, for him, is the only thing that has fundamental value. Much, of course, has non-fundamental value: pickaxes and dynamite, for example, which are reliable tools for mining for gold. Suppose now we ask the miner what he prefers: an ounce of gold, brought out of the ground with pickaxe and dynamite, or an ounce of gold, won by blind chance at the casino? If gold really is the only thing that has fundamental value for him, then he must be indifferent between these two treasures. Both are equivalent in terms of fundamental value; they only differ in non-fundamental value. If the miner prefers one of these prizes over the other, it is because he loves something other than gold.

Defenders of the cognitive achievement conception of knowledge (Sosa 2004; Greco 2010) have offered an answer to the Meno question. They argue that achievement is more valuable than (mere) success, and from this, along with the view that knowledge is an achievement and true belief a (mere) success, it follows that knowledge is more valuable than (mere) true belief. It is obscure whether the value of knowledge is epistemic value, on this view, but that is because the notion of "epistemic value" is obscure. Perhaps we should reject truth value monism: like a gold miner who values gold obtained through a hard day in the mine over gold won at the casino, perhaps we value truth acquired through intellectual virtue to truth acquired by luck. This will appeal to those already antipathetic to truth value monism, for example those who urge us to consider the epistemic value of understanding and wisdom, or those who urge us to consider not only the epistemic value of states such as knowledge and true belief, but also that of intellectual virtues such as open-mindedness, consistency, and intellectual integrity (see Zagzebski 1996).

A final possibility would be to reject the presupposition of the Meno question, and deny that knowledge is more valuable than (mere) true belief. If a

solution to the value problem eludes us, we must return to this presupposition and critically examine it. Whence the intuition that knowledge is more valuable than (mere) true belief? It is sometimes suggested (Kvanvig 2003) that a defense of the distinctive value of knowledge is needed to justify epistemologists' concern with knowledge. However, this argument is unsound. What might be relevant to justify epistemological inquiry would be the value of *knowing* about knowledge, not the value of knowledge. For there could easily be valuable knowledge about a worthless object: it is worth knowing about the molecular nature of rhodium, even if rhodium is worthless. (Moreover, it seems to us, no defense of the value of knowing about crustacean sexual behavior is needed to vindicate biological curiosity about it. Curiosity itself is vindication enough, we might say, although we do better to remember a liberal principle: no one needs to vindicate their preferred harmless (intellectual) activities, by showing that they are valuable.) The premise that epistemologists are wrong to study knowledge, unless knowledge is distinctively valuable, is false.

The Meno question concerns the value of knowledge, as opposed to true belief; a more basic question concerns the value of knowledge, as opposed to ignorance, in the form of a lack of belief, or in the form of false belief. Some philosophers defend the constitutive eudaimonic value of knowledge (Zagzebski 2003; Sosa 2003; Lynch 2004; Greco 2010; Baril 2010), others the instrumental eudaimonic value of knowledge (Kornblith 1993). It seems like some knowledge is worthless: I open the Wichita phonebook to a random page and memorize a name and phone number—an instance of knowledge, if ever there was one, but it is not plausible that this knowledge contributes to my well-being, either in an instrumental or a constitutive way. One might reply that such knowledge does have constitutive value, but that such value is trumped by something else, for example the time and effort it takes to memorize the number (Lynch 2004). But things do not *seem* that way: it is not like a case in which I must choose between pastrami and corned beef; even when I settle on pastrami, the appeal of the corned beef remains. In the case of the phonebook, knowledge seems to have *no* appeal.

Which knowledge has constitutive value? We might appeal to those truths that are objectively important or significant (Baril 2010), or to the answers to those questions that interest us or pique our curiosity. Perhaps we might be satisfied with our explanations coming to an end here: some items of knowledge have constitutive value, and others don't, and there's no further explanation of where the boundary between those species of knowledge lies.

An alternative to this would be to argue that knowledge is always *epistemically* valuable, even if instances of knowledge differ in their eudaimonic value. But this leads us back to the obscurity of the notion of "epistemic value." Truth value monism seems to imply that all instances of knowledge are

created equal, when it comes to their epistemic value. This would-be implication seems wrong: knowledge of particle physics seems not only eudaimonically but epistemically more valuable than knowledge of random phonebook entries. We can capture this intuition by saying something about “epistemic value”: epistemic value accrues when there is *mind-to-world fit*, that is when there is accurate representation of the world. Knowledge of particle physics involves greater mind-to-world fit than does knowledge of random phonebook entries—as Nick Treanor (forthcoming a) puts it, when I know the truths of physics I *know more* than when I know trivial truths. We might explain this by appeal to the idea that some true propositions contain more truth than others (Treanor forthcoming b), or to the idea that belief in some true propositions “carves nature at the joints,” while belief in others doesn’t (Sider 2011).

Suppose we were satisfied with some version of the claim that knowledge is valuable. What explains this? *Why* is knowledge valuable? Epistemologists sometimes appeal, in connection with the value of knowledge, to the first sentence of Aristotle’s *Metaphysics* on which “all men by nature desire knowledge.” Appealing to our curiosity, in connection with the value of knowledge, is puzzling, if one assumes *realism about the value of knowledge*, on which whether something is valuable is independent of whether anyone values it. If realism is true, that everyone desires knowledge neither entails, nor is entailed by, that knowledge is valuable. What Aristotle has in mind, of course, is that virtuous people desire knowledge, that people are curious if they are functioning properly, in accord with their human nature. Aristotle is no anti-realist about the value of knowledge; but he’s also not appealing to our mere curiosity in defense of the value of knowledge. And although in his famous sentence he says that everyone naturally wants knowledge (*to eidenai*), what he’s really defending (in the passages that follow) is an interest in the understanding of causes and principles (*episteme*) and in particular an interest in the understanding of fundamental causes and principles (*sophia*).

That said, perhaps *anti-realism about the value of knowledge*, on which the value of knowledge is in some sense exhausted by the fact that people value knowledge, is the way to go. Recently some philosophers have defended and criticized *epistemic expressivism* (see Chrisman forthcoming), on which evaluative or normative epistemic discourse should be understood as (among other things) expressing the speaker’s commitment to certain norms of belief formation. We might understand talk of the value of knowledge along similar lines.

Others have adopted a *hypothetical imperative approach* to the value of knowledge; on this view knowledge is good because it is a case of us successfully achieving something we desire (Papineau 1999; Dretske 2000; for criticism see Kelly 2003; Grimm 2008). This is a kind of realism (since it assumes the goodness of desire-satisfaction), but may not offend anti-realist intuitions.

Many epistemologists sympathetic with realism about epistemic value have turned to the idea that “belief aims at truth.” Where Aristotle saw a teleological link between people and knowledge; these contemporary thinkers see a teleological link between belief and truth. This might provide the explanation we seek: if “belief aims at truth,” then true belief (and therefore knowledge) is good, in as much as it is belief that has fulfilled its “aim.” Put another way, if “belief aims at truth,” then true belief (and therefore knowledge) is good, *qua* belief. But what does it mean to say that “belief aims at truth”? This metaphorical slogan must be given literal content. There are three possible approaches here. The first is a (relatively) *literal approach* that takes belief to require a psychologically real desire or intention (to believe the truth and nothing but the truth) on the part of any believer. This view is problematic in as much as this psychological claim is implausible. The second is a *Darwinian approach* that takes truth to be a biological function of belief; just as it is the biological “aim” of the heart to pump blood, it is the biological “aim” of belief to be true (Millikan 1984; Papineau 1993). This view is problematic in as much as it rests on empirically uninformed speculation about the natural history of cognition. The third is a *normative approach* that takes truth to be a “constitutive standard of correctness” for belief; on this view the goodness of true belief is a conceptual truth that flows *a priori* from the essential nature of belief (Wedgwood 2002; Shah 2003). This view is problematic in as much as it seems incompatible with philosophical naturalism; to escape this consequence, some of those who defend it favor epistemic expressivism (Shah and Velleman 2005).

Bibliography

- Ayer, A. J., 1956. *The Problem of Knowledge*. London: Macmillan.
- Baril, A., 2010. “A Eudaimonist Approach to the Problem of Significance.” *Acta Analytica*, 25, pp. 215–41.
- Buckwalter, W. and Stich, S., 2011. “Gender and the Philosophy Club.” *The Philosopher’s Magazine*, 52, pp. 60–5.
- Chrisman, M., Forthcoming. “Epistemic Expressivism.” *Philosophy Compass*.
- Cohen, S., 1999. “Contextualism, Skepticism, and the Structure of Reasons.” *Philosophical Perspectives 13: Epistemology*, pp. 57–89.
- , 1988. “How to Be a Fallibilist.” *Philosophical Perspectives*, 2, pp. 91–123.
- Conee, E. and Feldman, R., 2004. *Evidentialism: Essays in Epistemology*. Oxford: Clarendon.
- Craig, E., 1990. *Knowledge and the State of Nature*. Oxford: OUP.
- Cuneo, T., 2007. *The Normative Web*. Oxford: OUP.
- DeRose, K., 2009. *The Case for Contextualism: Knowledge, Skepticism and Context, Volume I*. Oxford: OUP.
- , 2004. “Single Scoreboard Semantics.” *Philosophical Studies*, 119, pp. 1–21.
- , 1995. “Solving the Skeptical Problem.” *Philosophical Review*, 104 (1), pp. 1–52.

- Dogramaci, S., Forthcoming. "Reverse Engineering Epistemic Evaluations." *Philosophy and Phenomenological Research*, 84(3), pp. 513–30.
- Dretske, F., 2000. "Norms, History, and the Constitution of the Mental." In *Perception, Knowledge, and Belief*. Cambridge: CUP, pp. 242–58.
- , 1970. "Epistemic Operators." *Journal of Philosophy*, 67, pp. 1007–23.
- Eng, B., 1990. "Is Realism Really the Best Hypothesis?" *Journal of Philosophy*, 87, pp. 667–8.
- Fantl, J. and McGrath, M., 2009. *Knowledge in an Uncertain World*. Oxford: OUP.
- Feldman, R., 2005. "Epistemological Puzzles about Disagreement." In Stephen Hetherington, ed., *Epistemology Futures*. Oxford: OUP, pp. 216–36.
- Field, H., 2009. "Epistemology without Metaphysics." *Philosophical Studies*, 143, pp. 249–90.
- Frances, B., Forthcoming. "Philosophical Renegades." In Jennifer Lackey and David Christensen, eds, *New Essays on Disagreement*. Oxford: OUP, pp. 121–66.
- , 2010. "The Reflective Epistemic Renegade." *Philosophy and Phenomenological Research*, 81, pp. 419–63.
- Fumerton, R., 1995. *Metaepistemology and Skepticism*. Rowman & Littlefield.
- Gettier, E., 1963. "Is Justified True Belief Knowledge?" *Analysis*, 23, pp. 121–3.
- Goldman, A., 1979. "What is Justified Belief?" In George S. Pappas, ed., *Justification and Knowledge*. Dordrecht: Reidel, pp. 1–23.
- , 1976. "Discrimination and perceptual knowledge," *Journal of Philosophy*, 73, pp. 771–91.
- , 1967. "A Causal Theory of Knowing." *Journal of Philosophy*, 64, pp. 357–72.
- Greco, J., 2010. *Achieving Knowledge*. Cambridge: CUP.
- Grimm, S., 2011. "On Intellectualism in Epistemology." *Mind*, 120, pp. 705–33.
- , 2008. "Epistemic Goals and Epistemic Values." *Philosophy and Phenomenological Research*, 77, pp. 725–44.
- Hawthorne, J., 2004. *Knowledge and Lotteries*. New York and Oxford: OUP.
- Hazlett, A., 2010. "The Myth of Factive Verbs." *Philosophy and Phenomenological Research*, 80, pp. 497–522.
- , 2009. "Knowledge and Conversation." *Philosophy and Phenomenological Research*, 78, pp. 591–620.
- Kelly, T., 2003. "Epistemic Rationality as Instrumental Rationality: A Critique." *Philosophy and Phenomenological Research*, 66, pp. 612–40.
- Kornblith, H., 2010. "Belief in the Face of Controversy." In Richard Feldman and Ted Warfield, eds, *Disagreement*. Oxford: OUP, pp. 29–52.
- , 2006. "Appeals to Intuition and the Ambitions of Epistemology." In S. Heatherington, ed., *Epistemology Futures*. Oxford: OUP, pp. 10–25.
- , 1993. "Epistemic Normativity." *Synthese* 94 (3), pp. 357–76.
- Kripke, S., 2011. "Nozick on Knowledge." In Saul Kripke, ed., *Philosophical Troubles. Collected Papers Vol I*. Oxford: OUP, pp. 162–224.
- Kusch, M., 2009. "Testimony and the Value of Knowledge." In A. Haddock, A. Millar, and D. Pritchard, eds, *Epistemic Value*. Oxford: OUP, pp. 60–94.
- Kvanvig, J., 2006. "Closure Principles." *Philosophy Compass*, 1, pp. 256–67.
- , 2003. *The Value of Knowledge and the Pursuit of Understanding*. Cambridge: CUP.
- Lewis, D., 1996. "Elusive Knowledge." *Australasian Journal of Philosophy*, 74, pp. 549–67.

- Lynch, M., 2004. *True to Life: Why Truth Matters*. MIT Press.
- MacFarlane, J., 2005. "The Assessment Sensitivity of Knowledge Attributions." *Oxford Studies in Epistemology*, 1, pp. 197–233.
- Millikan, R., 1984. *Language, Thought, and Other Biological Categories*. Boston, USA: MIT Press.
- Moore, G. E., 1939. "Proof of an External World." *Proceedings of the British Academy*, 25, pp. 273–300.
- Nagel, J., Forthcoming. "Intuitions and Experiments: A Defense of the Case Method in Epistemology." *Philosophy and Phenomenological Research*, 85, pp. 492–527.
- Nagel, T., 1979. "Moral Luck." In *Mortal Questions*. Cambridge: CUP, pp. 24–38.
- Nozick, R., 1981. *Philosophical Explanations*. Cambridge: Harvard University Press.
- Papineau, D., 1999. "Normativity and Judgment." *Proceedings of the Aristotelian Society*, 73, pp. 16–43.
- , 1993. *Philosophical Naturalism*. Blackwell.
- Planting, A., 1993. *Warrant and Proper Function*. Oxford: OUP.
- Pritchard, D., 2010. "Knowledge and Understanding." In A. Haddock, A. Millar, and D. Pritchard, *The Nature and Value of Knowledge: Three Investigations*. Oxford: OUP, pp. 66–90.
- , 2005. *Epistemic Luck*. Oxford: OUP.
- Pryor, J., 2000. "The Skeptic and the Dogmatist." *Noûs*, 34, pp. 517–49.
- Radford, C., 1966. "Knowledge—by Examples." *Analysis*, 27, 1–11.
- Ryle, G., 1949. *The Concept of Mind*. Chicago: University of Chicago Press.
- Rysiew, P., 2011. "Epistemic Contextualism." *The Stanford Encyclopedia of Philosophy* (Winter edn). E. N. Zalta, ed., Available at: <http://plato.stanford.edu/archives/win2011/entries/contextualism-epistemology/>
- , 2001. "The Context-Sensitivity of Knowledge Attributions." *Noûs*, 35, pp. 477–514.
- Sartwell, C., 1991. "Knowledge Is Merely True Belief." *American Philosophical Quarterly*, 28(2), pp. 157–65.
- Schaffer, J., 2007. "Knowing the Answer." *Philosophy and Phenomenological Research*, 75, pp. 383–403.
- , 2005. "Contrastive Knowledge." In Tamar Szabo Gendler and John Hawthorne, eds, *Oxford Studies in Epistemology* 1. Oxford: OUP, pp. 235–71.
- Shah, N., 2003. "How Truth Governs Belief." *Philosophical Review*, 112, pp. 447–82.
- Shah, N., and Velleman, D., 2005. "Doxastic Deliberation." *Philosophical Review*, 114, pp. 497–534.
- Sider, T., 2011. *Writing the Book of the World*. Oxford: OUP.
- Sosa, E., 2004. "Replies." In J. Greco, ed., *Ernest Sosa and His Critics*. Malden, MA: Blackwell Publishers, pp. 275–326.
- , 2003. "The Place of Truth in Epistemology." In M. DePaul and L. Zagzebski, eds, *Intellectual Virtue: Perspectives from Ethics and Epistemology*. Oxford: OUP, pp. 155–80.
- , 2000. "Skepticism and Contextualism." *Noûs*, 34, pp. 1–18.
- , 1999b. "How to Defeat Opposition to Moore." *Philosophical Perspectives*, 13, pp. 141–53.
- , 1999a. "How Must Knowledge Be Modally Related to What Is Known?" *Philosophical Topics*, 26, pp. 373–84.

- , 1991. *Knowledge in Perspective: Selected Essays in Epistemology*. Cambridge: CUP.
- Stanley, J., 2005. *Knowledge and Practical Interests*. New York and Oxford: OUP.
- Stroud, B., 1994. "Scepticism, 'Externalism', and the Goal of Epistemology." *Proceedings of the Aristotelian Society: Supplementary Volume*, LXVIII, pp. 291–307.
- Treanor, N., Forthcoming b. "Trivial Truths and the Aim of Inquiry." *Philosophy and Phenomenological Research*.
- , Forthcoming a. "The Measure of Knowledge." *Noûs*.
- Unger, P., 1975/2002. *Ignorance: The Case for Scepticism*. Oxford: OUP.
- Vogel, J., 1990. "Cartesian Skepticism and Inference to the Best Explanation." *Journal of Philosophy*, 87, 658–66.
- Wedgwood, R., 2002. "The Aim of Belief." *Philosophical Perspectives*, 16, pp. 267–97.
- Weinberg, J., Nichols, S., and Stich, S., 2001. "Normativity and Epistemic Intuitions." *Philosophical Topics*, 29, pp. 429–60.
- White, R., 2010. "You Just Believe That Because. . . ." *Philosophical Perspectives*, 24, pp. 573–615.
- Williams, B., 1973. "Deciding to Believe." In *Problems of the Self*. Cambridge: CUP, pp. 136–51.
- Williamson, T., 2000. *Knowledge and Its Limits*. Oxford: OUP.
- Wittgenstein, L., 1969. *On Certainty*, edited by G. E. M. Anscombe and G. H. von Wright, translated by G. E. M. Anscombe and D. Paul. Oxford: Blackwell.
- Zagzebski, L., 2003. "Intellectual Motivation and the Good of Truth." In M. DePaul and L. Zagzebski Linda, eds, *Intellectual Virtue: Perspectives from Ethics and Epistemology*. Oxford: OUP, pp. 135–54.
- , 1996. *Virtues of the Mind: An Inquiry Into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge: CUP.

27 The Philosophy of Perception: An Introduction

Paul Snowdon

It is a fact of central importance to our lives that we perceive, in a variety of ways, our environment and ourselves. According to conventional thinking we have five external senses, different modes of perception of the external world, and we also have internal channels of perception delivering information to us about ourselves. A concern with perception, especially external perception, has been part of philosophical thinking since the emergence of the subject.

1 The Importance of Perception

Concentrating initially on external perception, we can list various dependencies that reveal its importance. First, our knowledge of our environment, say that there is a tree ahead, entirely depends on our perception of the environment. Perception gives us, or at least plays a crucial role in giving us, such information in a way nothing else could. Now, although perception inherits in part its importance for us, and as a topic for philosophy, from its epistemological role, it is a working assumption of this account that there is a genuine distinction to be drawn between questions about perception and its nature and questions about knowledge and its nature. Second, our cognitive engagement with our environment rests on an ability to single out in thought, or attend to, objects around us. Thus, I see an object in the sky and wonder what it is. The question having been raised I can focus on the object and try to determine the answer. Further, given language, I can talk about that particular object with others. It seems obvious that these cognitive and communicative capacities, centrally involving attention, depend on our ability to *perceive* the items in question. Perception, then, is what enables us to do that. Third, we have to acquire the concepts we utilize in thinking about, and theorizing about, the world. Thus, I can form the view that there is a badger hiding in a

certain bush. To do this I need to have the concept or notion of a badger (and a bush). It seems obvious that the formation of such concepts depends on my perceptual encounters with my environment, more specifically, with badgers (and bushes). Putting this in a traditional way, human understanding depends on human perception.¹ Fourth, part of what we spend our lives doing and something to which we attach considerable value, is perceiving the environment; we listen to music, we watch sport, we look at pictures and nature, and we taste and savor food. It is not just that perception *aids* our cognitive lives directed at other things, exercising our perceptual capacities is a large part of our lives.

This manifest major importance of perception in our lives explains why all serious philosophers have engaged with it. When, as has sometimes happened, philosophers have neglected it, the philosophy they have produced has been impoverished and aetiolated.² Current philosophy is far from neglecting it.

2 The Philosophy of Perception

Although not all philosophers, at all times, who have discussed perception would put it like this, given our current perspective on philosophy we can divide the philosophy of perception into two (not completely equal) parts.

The first part starts from the observation that we have perceptual concepts, expressed in our perceptual language, for example the concept of seeing an object, or the concept of an object looking some way, or the concept of looking at an object, and so on. We can then attempt to spell out what follows from the truth of such claims. The goal is to spell out the entailments of such claims as; S saw object O. This is to attempt to do with perceptual concepts what philosophers have attempted to do with other concepts, such as those of knowledge, or action, or remembering, etc. This branch of the philosophy of perception can be described as the analysis of perceptual concepts.

Now, it is a matter of controversy how to informatively characterize what conceptual analysis is, but I shall assume in this introduction that we have an understanding of what it is and are not skeptical about the activity called "analysis." However, it does seem obvious that there is far more to a phenomenon than can be revealed by analyzing the concept that we possess of that phenomenon. This means that a proper theory of the phenomenon itself should not restrict itself to what, if anything, conceptual analysis can reveal.

So, the second, and undoubtedly the major part, of the philosophy of perception attempts to say, at a certain level, what is going on *when we do what we normally count as, say, seeing an object*. In current philosophy, this task is sometimes labeled "metaphysical" as opposed to the "conceptual" task. This

somewhat indirect way of posing the central issue is adopted to leave room for one attitude that is sometimes encountered in the philosophy of perception, which is that of regarding our ordinary or common sense way of categorizing such occurrences (as being genuine perceptions of external objects) is a mistake. A number of philosophers have adopted what J. L. Mackie usefully dubbed “error theories” about perception; that is theories according to which ordinary thought about such occurrences is in error.³

Now, given this specification of the central task of the nonconceptual component of the philosophy of perception two preliminary questions deserve to be highlighted immediately. One might be put in these words; is not the task specified above as the one for the philosophy of perception a task that a major branch of science is also dedicated to achieving? How then can philosophy itself make a contribution? This is, I believe, a genuine question that anyone engaged with the philosophy of perception, in its nonconceptual side, should keep before their minds. However, at this point, the best attitude to adopt in response to it is to engage open-mindedly with the proposals and arguments that philosophers of perception have advanced. That another and scientific discipline (or disciplines) is asking the same sort of question does not mean that the methods of philosophy have not got anywhere—which is, of course, not to say that they *have* got anywhere. The proof of this pudding, then, lies in the eating.

There is a second, and connected, preliminary remark that it is natural to voice. We are all familiar with the varieties of perception. Let us take as our example seeing. Now, we know that we are now seeing a page of a book, but if we simply try to scrutinize, as self-conscious subjects, the occurrence very carefully, it is remarkably hard to say anything informative about it. Reflective scrutiny, however intense, reveals little more than the presence of a spatially located page, with the features that are detectable, to sight—a result which is more or less trivial. In this fact lies a threat for that approach to the philosophy of perception carried out under the name of “phenomenology.” It aims to describe how it seems to people when they perceive. However sensitively this is done, why is it adding much to human understanding? Science is not so restricted, partly because scientists carry out real experiments, on the basis of which to make inferences, partly because they offer characterizations of aspects of perception at a lower level, that of the physical processes that are involved. How can philosophy (even when it is not purely phenomenological) generate informative results? The answer (or part of the answer) is that, despite not performing real physical or psychological experiments, it tries to locate evidence that supports in various ways unobvious theses. Again, the degree to which philosophers have done this will only emerge after looking at what they think of themselves as having unearthed.

3 The Sense-datum Theory

In introducing the philosophy of perception we need to start somewhere and the best place is by outlining and scrutinizing the theory that has traditionally been most popular among philosophers, (and also in earlier times, scientists) and that is what is called the “sense-datum” model. Starting at this point will not only familiarize readers with a major approach, certain arguments for it, some difficulties for it, but also equip them with some useful theoretical concepts.

3.1 The Theory

It would seem that when one sees an object one is having an experience, so the theory of perception is a branch of the theory of experience. It is in fact, and this is important, a branch of the philosophy of mind. Experiences seem to be a fundamental and very special sort of occurrence. The most popular way in current philosophy to characterize what is special about experience is the way that Nagel famously proposed—experiences are events that are like something for the subject to undergo.⁴ For present purposes I shall take this basic notion of an experience for granted, and assume that perception does involve experiences. In thinking about experiences, philosophers have found it tempting to analyze them in accordance with what has come to be called the act/object analysis. On such an analysis the basic structure of any such experiential occurrence is that there is an object or entity toward which a subject stands in a special mental relation, a relation that is called, when this model is expressed, the “act.” This structure is indeed the structure in terms of which we naively regard perceptual occurrences; they are regarded as a matter of a perceptible external thing coming within our experiential range. However, when this model is employed by philosophers in a completely general way in thinking about experience there is an important contrast between its standard significance in relation to perception, and its significance in relation to experience generally. The basic contrast is that the mental relation of perceiving understood normally is one that allows there to be a distortion of appearance even when the relation holds. However, when employed in the more general way the act or psychological relation is conceived of as “transparent,” meaning, perhaps among a number of things, that it reveals the encountered object precisely how it is. No question of illusions or misperceptions can arise.

Now, according to what is called the sense-datum theory, occurrences of experience, including the experiences involved in what we naively think of as perception of the world, consist of the obtaining of this transparent relation between a subject and an inner, mental, quality-bearing item. The best way to think of the theory is as postulating (at least in the case of vision) inner

mental pictures, which the subject directly apprehends. The inner item presents to the subject its sensible qualities. The whole occurrence amounts to an experience.

To fill out this traditional, but also modern, theory we need to specify the properties that the item is supposed to have, including what are conceived of as its conditions for existence. These properties are such features as real intrinsic color, generating lines and a two-dimensional spatial array, and being of a kind that exists only when perceived, that is, only when standing in such a relation to a subject. This is what Berkeley meant by calling the “esse” of ideas, which is what he called what was later called sense-data, “percipi,” that is to be perceived.

3.2 Arguments for the Theory

Why is this theory accepted? There are a number of familiar arguments in its favor, of which I shall present the two most popular, namely the argument from illusion and the argument from hallucination.⁵

Fundamentally, such arguments have a two-stage structure (as do many arguments in the philosophy of mind). They argue (or claim) that a certain conclusion applies to a range of cases, which we might call, for that argument, the base case, and then it is argued that if it is true for that range the same conclusion applied to all cases. The second claim can be called the generalizing or spreading step. The arguments tend to be named after the base cases they rely on—in the two I am presenting they are either illusions or hallucinations.

With the argument from illusion the base case are what philosophers call “illusions.” It does not matter whether they would be called “illusions” in the ordinary meaning of that expression. The defining feature of an illusion, in the philosophical use, and for the visual case, is that it is where an object that is seen looks a way it is not. A classic example is the straight stick that looks bent in water. How does the argument deal with this case? The best way to develop it, I propose, is to follow the exposition of an early master of this type of argument, namely Hume.⁶ His formulation of the argument is in terms of propositions containing demonstratives. Putting it in a series of propositions it runs:

- (a) That (whatever it is) is bent.
- (b) The external object in question is not bent (but straight).

Therefore:

- (c) That (which is bent) is not the external object.

Now, this is so far a negative conclusion, but if the premises are true it seems to be a valid application of Leibniz's Law about identity (that identicals must be indiscernible) to demonstrate a nonidentity, given a property difference. So far, nothing follows about the character and status of the item picked by "that" other than it is bent and not the external object. I think that we can represent what happens next in the argument by introducing a premise that says:

- (d) There is no physical thing that that item is identical with.

Therefore:

- (e) That is a nonphysical thing (which possesses bentness).

This claim seems to be very close to the conclusion about the case that is being aimed for, which is

- (f) That is a mental thing, that is a sense-datum.

The idea is that in illusion cases one can say that there is presented to the subject an item that he or she can focus on (in a style of thought which is demonstrative), and which is mental and possesses certain sensible qualities. Proponents of this conclusion say that standing in this relation to such an item can be called "being acquainted with it" or "directly perceiving it."

The interim conclusion just arrived at can be stated in a more general way; it is that in all cases of illusion the subject directly perceives a sense-datum, since the same pattern of argument can be provided for each case. The spreading step in this case can be expressed in the conditional claim:

- (g) Whatever is present in cases of illusions is present in all cases of apparently perceptual experience.

This is regarded as plausible in the light of such things as the similarity between illusory experiences and other kinds of experiences, and the fact that elements of illusion are more or less ubiquitous in experience. If that reasoning is accepted the final conclusion is reached:

- (h) In all apparently perceptual experience the subject directly perceives sense-data.

Now, this representation of the argument is to some extent arbitrary (in the way it divides up the premises), but it enables us to highlight the three main critical issues that can be, and have been, raised about it. Moving from the end to the start, these three things are: Is the spreading step well-supported? Does the argument move acceptably from the specific nonidentity thesis to the

introduction of sense-data in the illusion-involving case? Is the nonidentity thesis adequately supported?

It is the third question that I wish to focus on, but we can say about the first two that it is not completely obvious that the transitions they represent are correct. It would, perhaps, be best to regard them as plausible inferences, not necessitated by any compelling evidence. However, in this area that is the best we can achieve. What, though, of the third question, about the initial stages of the argument? That is, I believe, the point that deserves most attention. There is an obvious response to the argument. Why should we accept premise (a)? There is an alternative approach; we can substitute for (a) the following revised claim:

aR] that looks bent.

We can agree to that, without assenting to (a). Of course, if we stick with aR] then the argument for a nonidentity based on Leibniz's Law fails, since we can also agree that the stick, the relevant physical object, itself looks bent. There is no property difference on that account. Now we have arrived at a point where there is something like a principle in dispute. The principle has been called by Howard Robinson "The Phenomenal Principle," and it says that if it looks to a subject S as if there is something F, then there is something F.⁷ If we accept that principle then since the stick looks bent, something must be bent, and, presumably, it must be the object, whatever it is, picked out by "that." But although we can formulate this principle and doing so clarifies what is at issue, we remain without any reason to accept it. We can say instead: that certainly looks bent, but as to what is involved in something's looking bent I simply cannot say; all I can say is when something does look bent. It seems to me that this attitude is one that nothing can be brought forward to dislodge, and so the argument from illusion grinds to a halt. It has no entitlement to its treatment of the base case.

The so-called argument from hallucination utilizes hallucinations as its base cases. Now, the crucial difference between illusions and hallucinations is that it is built into the very notion of a hallucination that the experience does not consist in, or of, a perception of an external object. Indeed, the emergence into our conceptual scheme of the notion of hallucination, it is reasonable to suggest, derives from our realization that we can have experiences which, as undergone, seem to us to be perceptions, but which, on investigation cannot be counted as genuine perceptions. Once we realize that sort of thing happens we need the category of hallucination. This means that in an hallucination, if one were tempted to say "That is e.g., a physical object" then it is not true. In effect, claim (c) in the argument from illusion comes with our very conception of the base case *as* hallucinatory.⁸

This means that the traditional argument from hallucination, which as it were runs forward from that point along the same lines as the argument from illusion, faces the two other questions that the argument from illusion faced, namely; can it legitimately introduce sense-data as elements in the base case? And—can it spread the base case characterization to all cases? Now, there are, I believe, two things that stand out as problems for the traditional version of the argument. First, there is nothing other than a failure to imagine an alternative to the sense-datum analysis of experience in favor of introducing them in an account of hallucinations. Why must sense-data be involved in experiences of the hallucinatory kind?⁹ Second, even if we grant that sense-data are present in hallucinations, why spread that conclusion to other cases? The normal reply by exponents of this type of argument was that ordinary experiences are indistinguishable from hallucinations. To which Austin responded by pointing out that there is no compulsion at all to treat indistinguishable cases as being the same in nature. He used the example of looking at a lemon and looking at a bar of soap that looks exactly the same. They are indistinguishable, but do not have the same nature. There is no obvious reason to spread the characterization of the structure of the hallucinatory case to other cases. This traditional version of the argument, I suggest, fails. As we shall see though a more powerful version of an argument from hallucination has recently emerged.

3.3 Problems for Sense-data

I have argued that the traditional arguments do not provide much in the way of support for the sense-datum theory. They neither show that we need a uniform account of experience nor that if we do it should involve sense-data. This lack of support is striking. And indeed, since about the middle of the twentieth century there has been a general conviction that the theory is incorrect. I want to present two reasons in support of this opinion (without implying they are the only reasons). First, as I remarked earlier, the philosophy of perception is best thought of as a branch of the philosophy of mind. Perceiving is a psychological phenomenon. However, the dominant research strategy in the philosophy of mind has been physicalism—the idea that mental phenomena, including consciousness and cognition, have to be reducible in some sense to physical occurrences. This strategy is not without its difficulties, but it has been widely adopted. This means that it is a decisive problem for any theory if it cannot be integrated into that materialist framework. The difficulty for the sense-datum theory is that it contains two elements that are very difficult to fit into the materialist framework. First, there are the mental sense-data themselves. They are components in the experiential occurrence, but they do not seem to have a physical nature at all. They cannot be reduced to physical elements. Second, the psychological relation, the act, which is somehow a

pure apprehending of the object cannot be given a physical reduction either. No physical relation could have that nature. This means that the basic components of the theory defy materialist reduction.

The second reason flows from an apparent inconsistency between what we might call the dynamics of the sense-datum model and the phenomenon of perceptual experience itself. Now, although I have talked of the sense-datum model as a theory of perception there is a case for thinking that it is better described as a theory of visual appearance. One reason for saying this is that the model presupposes the relation between subject and object and it provides no account of it. Now, that relation has strong resemblances to the relation of perceiving. The idea is that when an object stands in that relation to a subject the subject is presented with that object, is acquainted with it. It seems to do for the subject what perceiving is thought of as doing. So it seems that the model presupposes something that is more or less a perceiving relation. What it does though with that relation embedded in the account is to explain why the subject has the appearances that he or she has. Roughly, when the subject apprehends a sense-datum that is genuinely red it will look to the subject as if there is something red. The model basically analyzes the occurrence of something such as its looking to S as if there is an F by regarding it as an F thing being apprehended. This reflects the central role of Robinson's phenomenal principle within the theory. However, there is a tension at this point between what the model seems to predict and how visual experience is. Consider the following case. Someone can, it seems, simultaneously see two lines that are some distance apart, and where we would say that neither line looks longer than the other, but they do not look to be the same length. They do not look to be the same length because the subject given all the time to scrutinize them has no inclination to hold that they are the same length. The subject would just say that he or she cannot tell how they are spatially related. We might summarize this by saying that the visual appearances are indeterminate. However, by contrast, within the model the lines on the sense-datum cannot be indeterminately related. Either they are the same length or different lengths. This seems to mean that according to the model there cannot be indeterminate appearances. There is, then, a case for saying that the dynamics of the model do not fit the nature of appearance.

The conclusion that seems indicated at this point is that it is unlikely that we want a theory of experience that incorporates the act/object structure. This reinforces the worry expressed earlier that traditional arguments for sense-data do not give reasons to introduce sense-data into the account. We have independent reasons to develop theories that leave them out.¹⁰ Now, there is a terminology that can be used to express this conclusion. At one time the idea of abandoning the act/object analysis of experience was put as the endorsement of what was known as an adverbial approach. The name comes from

the fact that in English we have sentences that have an act/object structure or grammar but where the phenomenon described by the sentence is more naturally thought of as not a relation to an object. A standard example is: S danced the Polka. On the surface this seems to be saying that S did something to an object picked out as “the polka.” But it is more natural to think of it as reporting the presence of an activity carried out in a certain manner. So an artificial but clearer re-expression would be: S danced in a polka manner. The role of the object noun is really adverbial. The idea of adverbialism in the philosophy of mind is that although experience reporting sentences can have an act/object structure, the real role of the noun is to characterize, in an adverbial manner, a process. Thus, “S felt a pain” is better thought of as saying something like “S felt in a pain-manner.” In reality such a proposal says little more than—do not think of the basic occurrence in an act/object way. So the general conclusion here can be stated as; adverbialism is preferable to sense-datum approaches. It is important not to conclude from this no experiences can have an act/object structure. Maybe genuine perceptual experiences have that structure. The conclusion is rather that act/object analyses are not obligatory.

4 Disjunctivism

I want to turn now to an approach that does treat ordinary perceptual experience as having an act/object structure, though without conceiving of the act in the totally transparent manner that is built into the sense-datum model.

It is striking that the most popular traditional philosophical model of perceiving, the sense-datum model, yields an account of perceptual experience that seems not to fit what it is like for the subject in various ways, for example, it not only seems to remove depth from visual appearance, but it also threatens to render inexplicable the cognitive role of perception, which is to yield knowledge of the external world. It threatens to do this by relocating the basic targets of the cognitive scrutiny that perception makes available to the subject within the subject’s mind, namely the private mental sense-data.¹¹ In an effort to preserve appearances and to sustain the idea that perceiving places us in cognitive contact with external objects, a different way of characterizing perception has emerged. This approach goes under the name of “disjunctivism.”

The view can be expounded in different ways, and so some of the details in the account developed here would not apply to all accounts that would claim to be versions of disjunctivism. The easiest way to approach the view is by locating an assumption about visual experience that philosophers often make. This has been called by Michael Hinton, who has had an important role in the emergence of disjunctivism, the assumption of a common visual element.¹² The assumption is that when we see things and when we hallucinate

the experience in each case is the same type of thing, with a shared intrinsic nature. There is, it is assumed, a common visual element. It is clear that sense-data theorists endorse this assumption, but non-sense-datum theorists can do so as well. If one makes this assumption the natural way to develop an account of the elements involved in genuinely seeing is to add to the experience a suitable environment to which it, the experience, is causally linked in the right way. Hallucinations are what results when the common visual element occurs outside the appropriate causally impinging environment. Disjunctivism rests on the idea that we should reject that assumption. In its place the account proposes that when one sees the environment that involves one type of experience, but when one hallucinates that involves another type or sort of experience. The difference in type or sort can be captured by saying that experiences of seeing have the (seen) object as a constituent, whereas experiences that are hallucinatory do not have such a constituent. The idea of two alternative forms of experience is sometimes expressed in a terminology that calls the perceptual cases the Good case and the hallucinatory case the Bad case (for an obvious reason). It can then be said that according to disjunctivism we can divide experiences into two classes, the Good class and the Bad class. The name of the view reflects the claim that experiences can be either Good cases or Bad cases.

What can be said in favor of this proposal? It is in fact not a very informative or explanatory characterization, but this is in line with a conviction that quite a few disjunctivists share, which is that we should not exaggerate the informativeness of which philosophy is capable. However, as a proposal it certainly fits the phenomenology of seeing. Our naïve sense is that we open our eyes and the object is present to us. We have no sense of an experience that is inner and beyond which the object stands producing it. That is quite an alien viewpoint. If the disjunctivist account is defensible it makes sense of how we view ordinary perception. Second, by regarding the experience as latching on to the object, it seems to fit the very fundamental idea that perception puts us in cognitive contact with the environment. Neither of these points is decisive, but it can be claimed that they earn disjunctivism the status of the default view.

There are two debates that seem to cut across disjunctivism, in the sense that one can take different views about them consistent with disjunctivism. The first issue, raised by McDowell's account of perception, is whether we should think of the experiences in visual experience as conceptual or non-conceptual.¹³ McDowell has suggested that the cognitive role of perceptual experiences requires that as states of experience they are concept-involving. But others have denied this. The second, and to some extent, related question is whether perceptual experiences should count as having content or not. McDowell defends the idea of content (which is conceptual) within a

disjunctivist approach, whereas other disjunctivists see no reason to do so.¹⁴ Both these issues suffer from it not being totally clear what the question is. Thus, for all its popularity in philosophy, talk of concepts remains obscure, as does talk of content.

I want, though, to engage with a central and crucial issue raised by disjunctivism. The denial by disjunctivism of the common element assumption has been challenged by a revised argument from hallucination, devised by, among others, Howard Robinson. The traditional argument from hallucination, which was discussed above, was meant to provide support for a sense-datum theory. Now, it did not do that, but it might be that an argument can be based on hallucinations in favor of the common element assumption. The new idea is to strengthen the spreading step in the argument by adding causal considerations, and to drop the disputable sense-datum idea when characterizing hallucinations. It can be said instead that in a hallucination the experience does not have an external object as a constituent. That is agreed. Now, the revised argument starts from the following idea. When we count as seeing our environment what happens in fact is that a causal process goes from the seen object along a route into the subject, via his or her eyes. Let us assume that NE1 is a neural event posterior to the ocular irradiation and caused by the irradiation, and which itself causes the experience involved in the perception. Now, it would seem reasonable to assume that if we could induce a neural event which is the same as NE1 by, say, direct stimulation of the subject's brain, it would cause the same sort of experience it causes in the successful case. This is simply an application of the principle expressible in the slogan—same cause, same effect. However, we can say that in the hallucinatory case the experience is inner and does not have any external object as its constituent. Putting all these assumptions together it seems to follow that since a cause like NE1 can cause a nonexternal object involving experience, that is the type of experiential effect it will always have, and hence the experience produced in a perceptual case will equally not be object involving.

Now, it is clear that this combination of assumptions is both in itself reasonable and yields a more cogent reason to generalize the characteristic of experience present in the hallucinatory case. Can the denial of the common element assumption, which is the core of disjunctivism, respond to this?

One radical response, developed by William Fish, is to deny that there are experiences in the hallucinatory case at all.¹⁵ What happens is that the subject is moved into a condition that he or she cannot tell is not a Good experience, but which in fact is no experience. This response faces two questions. The first is whether the denial of hallucinatory experiences is itself plausible. But the second is quite how maintaining this avoids the revised hallucinatory argument. It does deny an assumption that is made in the argument, but what is less than clear is where that denial leaves the argument overall.

An alternative way of trying to query this argument starts from the observation that a disjunctivist should not think of the experience in the normal case as caused by NE1. Rather, the experience must be regarded as having the external object as a constituent, and hence extending into the world. So, if producing NE1 (in a deviant way) does cause a hallucinatory experience it is not causing what is being produced in the Good case. That may be part of a response, but it remains to be made clear that it can avoid all the problems.¹⁶

This has only scratched the surface of a literature and approach that is being actively developed.

5 Intentionalism

Although disjunctivism is one popular approach, I need to introduce what seems to me to be the currently dominant view.

As a useful background to it we can look at the so-called belief theory of perception, one leading and ingenious proponent of which was George Pitcher.¹⁷ This theory developed out of two basic convictions. The first, which I have argued we should share, is that we need to avoid sense-data. The second is that in perceptual occurrences there is something that deserves to be called content. Thus, when S sees an item that item must look some way to S, a fact we can report by saying something of the form; it looks to S as if P (there is a bush). P should then be thought of as the content of that perceptual state. Now, what struck Pitcher, as a candidate theory that provides both these things is to characterize perception (at least initially) as the generation of beliefs about the environment. Nowadays no one would analyze belief in terms of sense-data, but beliefs have content—a belief has to be to the effect that P, for some value of P. This leads to a theory of perception according to which, in its simplest formulation, for S to see O is for there to be generated in S some beliefs about O via an eye-involving causal route. This treated seeing as the acquisition of a content-bearing psychological state, without involving sense-data. Ultimately, it was assumed of course, that belief would receive a materialist treatment.

There are some obvious problems with this proposal. One is that there is in fact a looseness between how things look and the beliefs that the subject will form. It can look to S as if P where the subject has reasons to think that not P. In such a case, S need not form a belief corresponding to the look. Having things look a certain way does not seem to involve believing things are that way. Further, it seems that we acquire beliefs all the time, without any looks corresponding to them. As S walks down the street, S no doubt is acquiring beliefs about the relative location of his starting point without it being the case

that these acquired beliefs correspond to looks. Third, it seems that the presence of a looks state involves what we might call an experience, whereas the acquisition of beliefs need not involve that. This seems to indicate that the content-bearing state should not be beliefs. Obviously, there are responses to be considered to these objections, but they seem to be worrying for the theory.¹⁸

This belief approach was replaced by a theory that is structurally similar, but instead of thinking of the content-bearing mental state acquired in perception as belief, in the new theory it is thought of as a distinct mental state bearing on contents. This can be expressed in various ways, but one way might be: one has an experience which represents it as true that *P*. In line with this people talk about the representational content that *P* of an experience. It is claimed that representational content need not lead to belief, that it amounts to an experience, and that its content can be prior to concepts.¹⁹ With these moves the objections to the belief theory are avoided. Since content-bearing states are also called by philosophers intentional states, one name for this approach is intentionalism.

Now, such theories can be divided, roughly, into two sorts. According to the first sort, the presence of a content-state of the right kind gives a total characterization of the experiential occurrence. According to the second sort, visual experiences have other features that contribute to the experience besides this sort of content. Peacocke, who belongs to the second group calls these further features the “sensational properties” of experience. There is a dispute over this within the intentionalist camp.

An advantage for the approach is that it accepts a common element assumption and so does not need to find a way out of arguments supporting that assumption. There are, though, serious issues that can be raised. Suppose that we ask: how does an experiences of a subject represent to the subject that *P*? What can the intentionalist theorist say? There is no obvious explanatory answer. Such theorists will probably simply say; they just do. This is linked to the fact that it is very hard to generate from this claim any testable consequences. What implications does it have to characterize perceptual experiences as having representational content? These consequences are hard to generate because there is no agreement as to whether the representational content requires, as some think, the subject to have concepts, or as others think, does not require concepts. It is also clear that the ascription does not imply anything about the physical processes that go on, since content-ascriptions have no obvious implications at that level. This absence of implications not only protects the theory from criticism, but also strips it of much content. Even as philosophical theories go, there is very little offered.

Further, we can put some pressure on the idea that propositional representational content can give an exhaustive account of the character of perceptual experience. If we are looking for structures that are in some way like visual

perceptions it is natural to consider pictures. I have already argued that visual perception does not involve pictures in the mind, but the tendency to think that way presumably reflects the sense that there is something alike between pictures and experiences. If the existence of such an analogy is accepted then we can provide a test of sort of representationalism, by considering whether we should think of pictures as describable in terms of propositional representational content. Suppose that we look at a picture such as Munch's *Scream*. We might say it represents there as being a person screaming. However, we cannot suppose that that captures the nature of the picture. We might add another proposition, say that it also represents that there is a path. What should strike us that however many represented propositions we add there will be something left out. Now, without more investigation and argument it would be incautious to conclude that we should not think of pictures this way, but it seems to be an incomplete approach. If there is an analogy here, then the same problem will exist for the theory of perception.

These are two worries, and the literature contains more.²⁰

6 Conceptual Issues

I have divided the philosophy of perception into two parts, one part labeled "metaphysical" (although that is a rather portentous nomenclature) and the other labeled "conceptual." In this section I wish to introduce some of the conceptual issues that have occasioned debate recently.

6.1 Uses of "see"

It is important when dealing with concepts to scrutinize our language. If we focus on the perceptual verb "to see," and indeed that is what most discussion has been about, one very fundamental distinction is often noted. On the one hand we couple that verb with expressions picking out, in a definite or indefinite way, objects, or entities, of various sorts. Thus we say, "S saw a tree" or "S sees President Obama" or "S saw the explosion." On the other hand we have a use where "see" is coupled with what philosophers call a proposition—such as—"S saw that it was raining" or "S sees that there is a tree." Now this second construction is sometimes called the factive construction, because it is generally agreed that we can only say with truth "S saw that P" where it is fact that P (i.e. only if it is true that P). It strikes us as inconsistent to say something like "S saw that the building was on fire but it was not on fire." We can instead say that "S thought he saw that the building was on fire but it was not," a remark that is not inconsistent. We can also make less committal reports using the word "looks," the term for

appearances of the visual sort. We can say “It looked to S as if the building was on fire,” a claim that can be true even if the building was not on fire. As well as being a factive construction “S saw that P” also seems to entail, or require, that S realized, that is to say, *knew*, that P. It reports a state of affairs that involves S knowing that the fact in question obtains. Thus, it is odd to say—S saw that P, but sadly he was not aware that P. This also seems to mean that it entails that S believed that P, given that knowledge (seems to) require belief. This use is sometimes called the “epistemic” us, but also the doxastic use of “see,” meaning it reports beliefs. It is also plausible to say that “S saw that P” not only entails both that P and that S knew (and believed) that P, but that it requires something like—S knew that P *in virtue of, or on the basis of, S’s seeing something*. There is a commitment to the knowledge accruing from an episode of *objectual seeing*.²¹ Now, not everyone will agree with these proposal but, roughly, these doxastic perceptual reports seem to represent our means of expressing the way we keep track of the gathering of information by using sight, achievements that are of central importance to us.

I want next to introduce some issues about the objectual use of “see.” First, in its main use the construction “S saw O” (where O is an entity) seems to require for its truth that O existed. In this “see” is like a verb such as “sat on.” S can sit on O only if O is there to be sat on, that is to say, O actually existed. But Anscombe famously pointed out that we do say such things as “S saw a bird in a nest” as way of reporting how things appeared to S, without committing ourselves to the actual presence of a bird in a nest. Her example to illustrate this was of an oculist saying to someone when looking at a complex screen—“Say when you see the bird in the nest”—where that means—when it looks as if there is a bird in the nest. Anscombe calls this the intentional use of “see,” and it seems that we do use “see” that way. The intentional use seems expressible in terms of “looks.” Thus, the oculist could have said; say when it looks as if there is a bird in the nest. However, the *nonintentional objectual* use is not a way of reporting appearances. Thus, if it is true that S saw a cat it hardly follows that it looked like a cat—maybe S just glimpsed the left ear or perhaps saw it in a dark place or perhaps it was disguised. The role of objectual seeing reports seems to be to say what objects happened to come into the subject’s view. It leaves quite open how much of the object came into his or her view, what the conditions for seeing were like, and what state the object in question was in. Such statements do not, it seems, convey a lot of information about the perceptual episode itself, aside from which items are involved in it.

But there are two important and much debated questions about the objectual use of “sees” which I wish to engage with. We can start by noting that it seems obvious that if S sees O then (a) S is having a visual experience, and (b) O exists. But what else is required?

6.2 A Differentiation Condition

One suggestion, proposed by Dretske, is

If S can see O then S can visually differentiate O from its immediate environment.²²

We could call this the “differentiation condition.” Whatever its precise content the structure of Dretske’s proposal is interesting. The relation of being able to differentiate links precisely the same two things that seeing links—the subject and the object. This proposal is attempting to spell out a relational element that is involved in the seeing relation. What is the differentiation relation? It means roughly that O must appear as different from its surroundings. Here is a case where that condition is not fulfilled. I paint a tank with a variety of green colors and place it in a greenish field. It does not know stand out in any way from its environment. It seems not at all unnatural to say that it *cannot* now be seen. So this negative judgment fits the differentiation condition. However, there are other cases and verdicts about them that seem harder to reconcile with the proposal. Suppose that I am looking through a hole cut in a white panel. Through the hole I can see, for example, a lawn. Now suppose that the hole is filled with a white plug that cannot be differentiated from the rest of the white panel. It seems that we should say that I am seeing the white plug that is obstructing my view of the lawn, even though that plug itself is not differentiable from the rest of the panel. These judgments give rise to the following questions; can the verdict in the white plug case be reconciled with Dretske’s proposal? If not, why is the disguised tank not seen? Do these verdicts bring out that the verb “see” is itself ambiguous?²³

6.3 A Causal Condition

It is, then, not entirely clear that Dretske’s differentiation condition is a necessary condition for seeing. However, there is a second condition for objectual seeing that it is widely believed has been brought out by conceptual analysis. That is the following condition, expressible in terms of the conditions already picked out. To conditions (a) and (b) must be added the causal condition: (c) the visual experience (reported in (a)) must be caused by the presence of the object O (reported in condition (b)).

This is one way of formulating the central claim in what is called the Causal Theory of Seeing. Many people endorse this claim about seeing, but many also endorse similar causal conditions for the application of lots of our psychological notions or concepts, such as acting, knowing, and remembering.

In considering this proposal as a piece of conceptual analysis we must ignore the idea that a causal condition might be proposed as a necessary

condition on empirical grounds. The present proposal is that we can unearth *conceptual reasons* to regard it as necessary for seeing. What reasons are they? A number have been advanced but the crucial and most influential reason is one made famous by H. P. Grice. Grice considered a case such as the following. Suppose S is standing in front of a clock. S is having a visual experience as of a clock. This amounts to supposing that versions of conditions (a) and (b) obtain. What is happening though is that an evil scientist is giving S an hallucination as of a clock, by manipulating S's visual cortex. Now, it would be agreed, surely, that S is not seeing the actual clock, despite the match between environment and visual experience.

If we agree with that judgment the question that arises is; why is the external clock *not seen*? The hypothesis that Grice proposes is that it is not seen because it is a necessary condition for an object to be seen that it causes the visual experience. So the correctness of (c) is regarded as the best explanation of the truth of the agreed negative judgment.

If that conclusion is drawn it is obvious that the analysis is not complete. This is obvious because someone can be having a visual experience, and there be X in the environment, and the experience causally depends on X, and yet X is not seen. Thus, in all visual experience the brain of the subject fulfills these conditions. Or again, it might be that the evil scientist in the above case was caused to produce the clock hallucination because there was a clock in front of S. The acknowledgment of such cases generates the problem that became known as that of deviant causal chains. Some cases of causal dependency of experience on objects are deviant, in the sense that they do not amount to sightings of the object. The task is to characterize the difference between deviant and no-deviant causal chains. Massive ingenuity has been devoted to solving this problem within the Causal Theory approach.

But the question can also be raised as to whether the pro-causal argument is convincing. This question starts from the recognition it is assumed in the presentation of the argument that the visual experience the subject is one thing and the external object another. Putting it in a slightly different way the argument is developed in the context of just assuming something like the common visual element. The issue is how they can be related? Now we have already encountered the disjunctivist conception of perception, and considered whether it might be a correct account of perception. According to it, in the perceptual case, the subject has an experience of a type in which the perceived object is a constituent. Now, the chief difficulty with the view is that there is a quite persuasive but broadly empirical argument in favor of what was called the common element hypothesis. Whatever the power of that argument, its character means that it does not rule out the possibility of a disjunctivist understanding of our basic perceptual *concepts*. This opens up to us an alternative explanation of the negative judgment about the hallucination case

that figured in Grice's type of argument. We can say that the case as described is not a perception of the actual external clock because in order to be a perception the experience in the case must be one in which that external clock was a constituent, but the described origin of the experience, being induced by an evil scientist, means that it could not be an experience in which the clock was a constituent. The evil scientist could have induced it whether there was a clock there or not. The role in relation to this type of argument of the disjunctivist idea is to provide a potential alternative explanation (not involving the idea that perception has to be causal) for the data which the original argument assumes supports the postulation of a causal condition. In this way disjunctivism illustrates how an idea or proposal can relate to different issues, and it may be differently effective in different cases.

At this point in the argument, at least three questions arise. The first, which I shall not pursue, is whether this alternative explanation is really a cogent alternative to the causal one. The second, is whether there might not be other reasons for adopting the Causal Analysis, which this disjunctivist approach does not undermine. The third is what sort of analysis or account of our perceptual concepts is possible if we do not immediately go in the causal direction?²⁴

In relation to the third question, one possible account is that our conceptual fix on seeing is as a distinctive relation that can obtain between a subject and an object, which has a certain cognitive functional role for us. It is a relation which when it obtains enables us, given our cognitive capacities, to pick the object out (although not necessarily in virtue of discriminating it). The precise nature of that relation is to be fixed by empirical investigation. This account seems to fit the way we know when we see. It is not that we can detect the seeing itself. Rather we focus on the world—on the seen object—and infer that we are seeing it. The proposal is that we can illuminate many aspects of our thought and knowledge about seeing by approaching the concept in some such way.

These conceptual and linguistic investigations are speculative and conjectural, but they have formed a significant part of recent philosophy of perception.

7 Conclusion

In a short introductory essay it is impossible to introduce into the discussion all the ideas that are worth considering. It may very well be felt, and indeed be true, that among the ideas I have outlined there is no properly satisfactory model of perception. But the next stage should be to read more generally in the literature that this fascinating and central phenomenon has elicited from philosophers.

Notes

I wish to thank Howard and Barry for the invitation to contribute to this volume and their encouragement during the writing stage. I also wish to thank Craig French for recent conversations, and Mike Martin and Mark Kalderon for earlier conversations, which have strongly influenced my approach.

- 1 Plotting this dependence is a central task of Locke's *Essay Concerning Human Understanding*, and by talking of the understanding I am deliberately alluding to Locke's program.
- 2 I have in mind the period in the 1970s and 1980s when philosophers under the influence of Davidson whose basic ideas about thought and language ignored perception and experience, and of Putnam whose powerful arguments in favor of functionalism lead philosophers to replace talk of perception with talk of causal connexions.
- 3 See Mackie 1973, p. 45 for an explanation of this very useful notion.
- 4 See Nagel 1974. There is a dissenting tradition—for one version of that see Snowdon 2010.
- 5 It is important to be careful when coming across what are called by their authors arguments from illusion and hallucination. The name merely marks the fact that these phenomena figure as elements in the arguments. It does not tell us how they figure. Not all arguments called "arguments from illusion" are aiming at the same conclusion.
- 6 See Hume 1982.
- 7 See Robinson 1997, ch. 2.
- 8 To say that the claim (c) is present, as one might say, for free, should not be taken to imply that there is an internal item that the demonstrative actually picks out. What comes for free is simply the incorrectness of the positive demonstrative claim.
- 9 This failure can be expressed as the failure to note the possibility of an adverbial approach, the notion of which is explained later in the discussion of sense-data.
- 10 A very good critical discussion of the sense-datum theory and its arguments can be found in Pitcher 1971 and a more sympathetic discussion in Robinson 1997. The classic criticism of it is very entertainingly given in Austin 1962. Smith 2002 is also a rich and interesting recent discussion.
- 11 The idea that we should shape our account of perception to explain the knowledge it yields us is strongly developed by John McDowell, see McDowell 1982.
- 12 See Hinton 1973, sec. IIb.
- 13 See McDowell 1982.
- 14 To pursue these issues see McDowell 1996 and Travis 2004.
- 15 See Fish 2009. Fish's book is a good introduction to disjunctivism.
- 16 A very good place to read further into the debate about disjunctivism is (eds) Byrne and Logue 2009.
- 17 See the final part of Pitcher 1971.
- 18 Another issue that can be raised is that we tend to think that beliefs are mental states that are secondary to experience and perception. Beliefs require concepts, which need acquiring, and therefore they seem to be posterior to perception.
- 19 One very effective exposition of such a view is Peacocke 1983.
- 20 Attempts to argue against intentionalism can be found in Campbell 2002, Martin 2002, and Travis 2004.
- 21 Anyone wishing to follow up these linguistic claims and issues can find detailed discussion in Dretske 1979, Cassam 2007, and Breckendridge forthcoming.
- 22 See Dretske 1969, ch. 2.
- 23 To follow up this debate, Dretske 1969, ch. 2 and Cassam 2007, ch. 3, are good places to start.
- 24 This debate about the causal condition can be pursued by reading Grice 1961, Snowdon 1980, Child 1994, and some of the relevant essays in Roessler et al. 2011.

Bibliography

- Anscombe, G. E., 1965. "The Intentionality of Sensation." In R. J. Butler, ed., *Analytical Philosophy, Second Series*. Oxford: Blackwell, pp. 158–80.
- Austin, J. L., 1962. *Sense and Sensibilia*. Oxford: OUP.
- Breckenridge, W., forthcoming. *Visual Experiences: A Semantic Approach*. Oxford: OUP.
- Byrne, A. and Logue, H., eds, 2009. *Disjunctivism*. London: MIT Press.
- Campbell, J., 2002. *Reference and Consciousness*. Oxford: OUP.
- Cassam, Q., 2007. *The Possibility of Knowledge*. Oxford: OUP.
- Child, W., 1994. *Causality, Interpretation and the Mind*. Oxford: OUP.
- Crane, T., ed., 1992. *The Contents of Experience*. Cambridge: CUP.
- Dancy, J., 1988. *Perceptual Knowledge*. Oxford: OUP.
- Dretske, F., 1969. *Seeing and Knowing*. London: Routledge and Kegan Paul.
- Fish, W., 2009. *Perception, Hallucination and Illusion*. Oxford: OUP.
- Grice, H. P., 1961. "A Causal Theory of Perception." *Proceedings of the Aristotelian Society*, 35, pp. 121–52; the crucial sections are reprinted in Dancy 1988.
- Hinton, M., 1973. *Experiences*. Oxford: OUP.
- Hume, D., 1982. *Inquiry Concerning Human Understanding*. Oxford: Clarendon.
- McDowell, J., 1996. *Mind and World*. Cambridge, MA: Harvard University Press.
- , 1982. "Criteria, Defeasibility and Knowledge." *Proceedings of the British Academy*, 68, pp. 455–79; the crucial sections are reprinted in Dancy 1988.
- Mackie, J. L., 1977. *Ethics*. London: Penguin.
- Martin, M., 2002. "The Transparency of Experience." *Mind and Language*, 17, pp. 376–425.
- Nagel, T., 1979. *Mortal Questions*. Cambridge: CUP.
- , 1974. "What Is It Like to Be a Bat?" Reprinted in Nagel 1979.
- Peacocke, C., 1983. *Sense and Content*. Oxford: OUP.
- Pitcher, G., 1971. *A Theory of Perception*. Princeton: Princeton University Press.
- Roessler, J., Lerman, H., and Eilan, N., eds, 2011. *Perception, Causation and Objectivity*. Oxford: OUP.
- Robinson, H., 1994. *Perception*. London: RKPP.
- Russell, B., 1912. *Problems of Philosophy*. Oxford: OUP.
- Smith, D., 2002. *The Problem of Perception*. London: Harvard University Press.
- Snowdon, P. F., 2010. "On the 'What-It-Is-Likeness' of Experience." *The Southern Journal of Philosophy*, 48, pp. 8–27.
- , 1992. "How to Interpret 'Direct Perception.'" In T. Crane, ed., pp. 48–78.
- , 1980. "Perception, Vision and Causation." *The Proceedings of the Aristotelian Society Supplementary Volume*. Reprinted in Dancy 1988.
- Travis, C., 2004. "The Silence of the Senses." *Mind*, 113, pp. 57–94.

28 Practical Reasons: The Problem of Gridlock

Ruth Chang

Philosophical ethics in the last century was occupied by two main lines of investigation: *ethical theorizing*—clarifications of and refinements to theories of morally right action, primarily consequentialism, deontology, and virtue ethics—and, *meta-ethics*—the application of ideas from metaphysics, epistemology, and the philosophy of language to the nature of moral thought and discourse. In this century, at least so far, much of the most exciting work in philosophical ethics turns away from these mainstays and focuses on a more general and deeper set of issues that straddle and go beyond ethical theory and meta-ethics, what we might call *the philosophy of practical reason*.

The philosophy of practical reason ranges over a number of issues but its most central concern the nature of normative practical reasons—the considerations that support or count in favor of performing some action or having some attitude. Normative practical reasons include moral reasons but many more besides. They are arguably the most basic building blocks of theorizing in any normative domain and the most basic subject matter of meta-normative theorizing.¹

The chapter has two aims. The first is to propose a general framework for organizing some central questions about normative practical reasons in a way that separates importantly distinct issues that are easily run together. Setting out this framework provides a snapshot of the leading types of view about practical reasons as well as a deeper understanding of what are widely regarded to be some of their most serious difficulties.

The second aim is to use the proposed framework to uncover and diagnose is a structural problem that plagues the debate about practical reasons. A striking feature of this debate is that it has been marked by the persistence of three dominant types of view. The problem is not that the same types of view persist—that, alas, might be a feature of the philosophical condition—but *why* these particular types do. As I will suggest, these types persist because they make substantive assumptions in answer to one question of the framework, which in turn have profound effects on how arguments for and against one

type of view relate to arguments for and against another type. In short, arguments favoring one type of view have merit largely only given that substantive assumption, while arguments against it have force largely only given a different substantive assumption. As a result, a common move in the debate involves a proponent of one type of view offering what she and others proposing that type consider to be a devastating criticism of an opposing type of view, only to find that her criticism is shrugged off by her opponents as easy to answer, misguided, or having little significance for their view.² This isn't, I will suggest, due to conceptual blindness or mere slavish devotion to a theory but something fundamental about the argumentative structure of a debate over genuinely shared issues. Hence, the debate about practical reasons suffers from argumentative gridlock. The proposed framework helps us to see why this is so, and, as I will tentatively suggest at the end of the paper, what we might do to move beyond it.

1 A Framework

1.1 Three Questions about Normative Practical Reasons

Debate about practical reasons might be usefully organized around three meta-normative questions:

- (a) What is the *content* of normative practical reasons?
- (b) What is the *nature* of their normative force?
- (c) What is the *source* of this normative force?

These questions are “meta-normative” in that they are, on their face, meta-physical questions about the general nature of practical reasons and not substantive, normative questions about what reasons we have or the substantive conditions under which, by the lights of a normative theory, we have them.³ As we will see, these three questions provide a concrete set of issues by which we can gain both a synoptic view of the leading theories about practical reasons and a deeper understanding of what might be regarded as their main difficulties.

First. Which sorts of considerations—let's assume that they are facts—are normative practical reasons? On the face of it, a wide variety of facts can be a normative reason for you to do something—that you promised to, that you want to, that doing it would be good in some way, and so on. But perhaps these facts can be systematized so that some are derivative while others not, with all the nonderivative reasons being of a single unified sort. Is there a single type of fact that systematically carries the action-guidingness of a reason?

Philosophers have offered three broad answers to this question. “Desire-based” reasons theorists think that when we systematize our reasons, we will see that our practical reasons are at bottom facts that we want or would want—under certain evaluatively neutral conditions—something.⁴ “Value-based” theorists, by contrast, maintain that our practical reasons are given by evaluative facts about the object of our desires or, according to “buck-passers” about value, by the nonevaluative facts that subvene these evaluative facts.⁵ So, according to desire-based views, your reason to have some ice cream is given by the fact that you want some or would want some under evaluatively neutral conditions, while according to value-based views, it is given by facts about the object of your desire, such as that the ice cream is creamy, delicious, or would give you pleasure.

A third answer is hybrid or pluralist: both sorts of considerations are needed systematically to account for our reasons—not all reasons can be facts about the goodness of something (or the facts that subvene those facts) and not all reasons can be facts that someone wants something or would want it under certain neutral conditions. If we try to systematize our reasons, we will find that at bottom, some of our reasons are facts that we want things while others are facts about the goodness or other features of what we want.⁶

Views about which sorts of facts must figure in a systematic account of our reasons are views concerning which kinds of fact are most fundamentally and nonderivatively our reasons. We might loosely say that these facts give the “content” of practical reasons. Sometimes the debate about which considerations systematically bear the normativity of a reason is run together with issues concerning source, since the questions are often not clearly distinguished. We’ll have more to say about the source question below.

Second. What is it to be a practical reason? Reasons are normative—they are action- and attitude-guiding—but what is it to be action-guiding? Or put another way, reasons “count in favor” of action or attitudes, but what is the nature of counting in favor?

Again, there are three main answers. Some philosophers assume that normativity is a motivating force, that is, a psychological force that causes or psychologically compels a “rational” agent—that is, an agent who meets certain evaluatively neutral conditions—to act.⁷ A reason counts in favor of action just in case a structurally rational agent who recognizes it is thereby psychologically moved to perform the action; it guides his/her action by motivating him/her to do it. What it is to be a reason is to be a consideration that motivates an agent who satisfies certain evaluatively neutral conditions to action.

Others treat normativity as a *sui generis* justificatory force. A reason counts in favor of action in that it *pro tanto* justifies or supports that action.⁸ When a rational agent recognizes a reason, her action—and her motivation—is guided by the justification for performing it. What it is to be a reason is to be a

consideration that justifies or supports performing an action or having an attitude.⁹ The way a reason guides an agent's action according to this view is different from the way it motivates him/her; it guides his/her action by pro tanto justifying it, and pro tanto motivation then follows in the "rational" agent who is motivated to do what he/she recognizes he/she has a reason to do.

A third idea tries to have it both ways by suggesting that practical normativity is some kind of "volitional" force and is thereby both motivational and authoritative (and thus justificatory) for the agent.¹⁰ A reason counts in favor by having a special volitional authority for the "rational" agent, that is, an agent whose will, as a constitutive matter, obeys the laws of practical reason. When faced with a reason, a rational agent is volitionally moved to act in accordance with that reason, and that reason is authoritative for that agent.¹¹ What it is to be a reason, then, is to be a consideration that volitionally compels an agent whose will conforms to the laws of practical reason to action.

Views about what it is to be a reason, that is, to "count in favor of" an action, are views concerning the *nature* of normative force.

Third. Where does the normativity of a reason come from? The question of the *content* of a practical reason is, "What sort of consideration systematically bears the action-guidingness of a practical reason?" The question of the *nature* of normative force is, "What kind of action-guidingness does a practical reason have?" But we can also ask a further question: where does the normativity of a practical reason come from—whatever sort of fact it might be and whatever kind of normative force it might have? Holding the *content* of a practical reason and the *nature* of its normative force fixed, we can ask, what is the *source* of a reason's normativity?

Note that the question of source is not the substantive question: "Under what substantive *conditions* is a consideration a reason according to a normative theory or principle?" That is, according to a substantive normative theory—such as consequentialism or deontology—what *conditions* must obtain in order for a fact to be a reason? Some possible answers might be: "When the consideration in those circumstances would make the action maximize happiness for the greatest number" or "When it would render the action the fulfillment of one's moral duty," and so on. This is a *normative*, not a metaphysical, question about practical reasons and thus is no part of our meta-normative framework.¹²

1.2 Normative Source

While the questions of *content* and *nature* are relatively straightforward, the question of the *source* requires further explanation. Where does the normativity of a reason come from?¹³ There are multiple questions here that need to be distinguished. One is, "What is the *cause* of something's having the normativity of a reason?" This is not the question of source. In wanting to know

the source of normativity, we don't mean to ask what causes something to be normative, if indeed that question makes sense. Instead, we are looking for something deeper—roughly, what *metaphysically determines* something's being a reason.

Another question, somewhat closer to what we have in mind, asks what is the *modally subvening base* of something's being a reason, that is, which facts modally covary with the fact that something has the action-guidingness of a reason. The source of normativity, however, isn't what modally covaries with something's being a reason—it isn't the fact, for instance, that must change when there is a change in the fact that something is a reason. Consider an analogy. Suppose God is the source of morality; God—or a supernatural realm—is where morality comes from. This view about source is consistent with the idea that moral facts supervene on natural facts, that you can't have a change in the moral facts without having a change in the natural facts. So the sense of "source" being sought here is not that of a modally subvening base.

A third question asks what general principle or law *subsumes* the fact that something is a reason. There is a sense in which the normativity of specific reasons may "come from" more general normative reasons or principles, if the most extreme forms of particularism are false. A specific reason may be an instance or instantiation of a general principle such that when it holds, it explains why a general principle holds. But when we wonder about the source of normativity, our question is not about the subsumption of particular reasons under general normative reasons but rather about one thing being the metaphysical fount of another. When we ask for the source of normativity by asking "Where does the normativity of a reason come from?" we are looking for an explanatory connection that holds of metaphysical necessity but is neither cause, supervenience, nor subsumption.

Intuitively when we ask about the source of normativity, we are asking what "makes" a fact normative, what it is "in virtue of which" some fact has the normativity of a practical reason. If we trace the normativity of a reason back to its fount, we will reach what "makes" the consideration a reason in the first place—its normative source. As metaphysicians might say, we are looking for the metaphysical *ground* of something's being a reason.¹⁴ There is burgeoning literature on "ground" we needn't engage; for our purposes, we can work with the basic idea that *x grounds y* when *x* gives a metaphysically necessary explanation of *y* that is not causal, modal covariation, or subsumption.

Now there is more than one way in which one fact can make something the case, different ways in which a fact can be grounded. The most natural way one fact can ground another is by *constituting* it. The fact that *p* or the fact that *q* ground the fact that *p* or *q* in that the former facts constitute the latter fact. The fact that it's H₂O grounds the fact that it's water in that being H₂O constitutes being water. Or consider causation. Striking a match causes it to light. What "makes" the striking cause the lighting, that is, what constitutes the fact

that the striking causes the lighting? One answer might be “a nomological law according to which under conditions C, a striking of a match causes the match to light,” and another might be “a set of regularities whereby a striking of a match under conditions C is followed by its being lit.” This law, or this regularity, is what constitutes the fact that the striking causes the lighting. It is where the causality of the striking comes from. Tracing the causality back to its fount by discovering what constitutes something’s having it, we end at the law or regularity.

So one way to answer the question, What is the source of x? is by saying what constitutes the fact that x. And thus one way of answering the source question about normativity is by saying what constitutes the fact that something is a reason.

Another way something can—perhaps degenerately—be grounded is by being “self-grounded,” that is, by being its own fount. God may be *causa sui*, the cause or ground of himself. Or consider, again, the case of cause. If we ask, “Where does a law that constitutes the fact that one event causes another come from?,” the answer may be “Nowhere” or, for our purposes, what we can take to be its equivalent, “From itself.” The law is a nomological necessity. All necessary truths are self-grounded; there is nothing further that explains them. Contingent brute facts can also be self-grounded. Suppose we ask, “Where does the negative charge of an electron come from?” The answer might be “From the fact that an electron has a negative charge”; there’s no more explanation to be had, end of story. Facts that are explanatorily primitive are self-grounded; they cannot be accounted for in any other terms—they represent the end of the line in explanation—and hence are their own ground.

So another way to answer the source question is by appealing to the fact whose source we are seeking in the first place—the fact is its own source. If we ask where the normativity of a reason comes from, one possible answer is that there is nothing further that explains why a fact has the normativity of a reason other than the fact that it has the normativity of a reason. The fact that something is a reason might be its own normative source.

I believe there is a third way in which one fact can intuitively “make” or “ground” another that is neither a case of constitution nor one of self-grounding. This is a relation of *metaphysical creation*. Consider, again, the case of cause. The fact that the striking of the match causes the lighting may be constituted by nomological laws, but what is the source of those nomological laws? Where do they come from? We might answer the source question by appealing to God who makes those laws. God is the source of the nomological laws in that he metaphysically creates them, where creation is a kind of metaphysical determination beyond causation or constitution. (More needs to be said about this relation which cannot be said here).¹⁵ So a final possible way to answer the question, “What is the source of x?” is to say what metaphysically creates x—whatever that turns out to be.

There are thus three ways in which the question, “what makes something a reason for action?” can be answered. The fact that something is a reason can be what constitutes the fact that it is a reason; it can (degenerately) be its own ground or source and finally, the source of a reason can be what metaphysically creates the fact that it is a reason.

As might be expected, philosophers can be seen as having offered three main answers to the source question, each of which broadly corresponds to one of the three ways the source question can be answered. “Source externalists” think that normative facts make some fact, like the fact that it is painful, is a reason. So the source of normativity is external normative facts. These facts are “external” in the sense that they lie outside of us as agents.

Some source externalists think that the normative facts that ground the fact that something is a reason are those very facts; when we ask what “makes” something a reason, our answer is “nothing,” or, equivalently for our purposes, “the fact itself.” In this way, the fact that something is a reason is self-grounded. Put another way, when we contemplate the fact that something is a reason, we are already at the source of the normativity of that reason. Other source externalists think that the normative facts that ground the fact that something is a reason are *other* normative facts, facts not about reasons but about values, for example, evaluative facts about the goodness of things. The constitutive ground of the fact that being painful is a reason to avoid it is the badness of the experience, and hence it is the disvalue of the experience that is the source of the reason’s normativity.¹⁶ So the source of the normativity of your reason — “it’s painful!” — to avoid touching the hot poker is either the fact that its being painful is such a reason or the fact that pain is bad.¹⁷

While normative externalists can be said to locate the source of normativity outside of us, in a realm of normative facts, “normative internalists” think that normativity has its source inside of us, and in particular, in desires and dispositions—the mental states toward which we are largely passive. If the fact that an experience is painful gives you a reason to avoid it, it does so in virtue of the fact that you want—or would want under certain evaluatively neutral conditions—to avoid pain. What constitutes the fact that something is a reason is thus some relation between that thing and one’s desires or dispositions. One way something might relate to your desires is by being constitutive of its satisfaction. Suppose you want pleasurable experiences. What constitutes the fact that being pleasurable is a reason for you to pursue it? The fact that you want pleasure and that being pleasurable constitutes satisfaction of that desire. Another way something can relate to your desires is by being instrumental to its satisfaction. What constitutes the fact that being painful is a reason to avoid it? The fact that you want to concentrate on writing your paper, and the pain would be distracting.¹⁸

Source externalism and internalism occupy the bulk of discourse about the source of normativity. Each appeals to one or other of the first two explanatory connections of grounding—self-grounding and constitution. Together they offer up a neat dichotomy in thinking about normative source—it is grounded either in facts external to us or in our internal dispositions, desires, and motivations.

There is, however, a third view, what we might call “source voluntarism.” According to voluntarism, normativity comes from an act of will. Like internalism, voluntarism locates the source of normativity inside of us—but not in passive states like desiring but rather in the *active* state of willing. Divine command theory offers the earliest example of such a view; by willing it, God can ground the fact that being a hoofed animal is a reason not to eat it. Post-enlightenment, philosophers replaced God’s will with our own; through an act of will, a rational agent can lay down laws for himself/herself. A rational agent’s own legislation can ground the fact that something is a reason. Kant’s revolutionary account of normativity is, on some interpretations, the most developed defense we have of voluntarism, but others before him—Hobbes, Locke, and Pufendorf—arguably helped to lay the groundwork for such a view.¹⁹

An interesting feature of source voluntarism is that, unlike both source externalism and source internalism, voluntarism can in principle provide an answer to the source question via either the relation of constitution or the relation of metaphysical creation. An act of will can be the constitutive ground of something’s being a reason and it can also be what creates, as opposed to constitutes, the fact that it is a reason. If, however, the will is to be a source of normativity, it is more plausibly via metaphysical creation rather than constitution, for how can an activity, such as willing, constitute the fact that something is a reason? Indeed, if there is a relation of metaphysical creation, then willing is a most natural relatum. It’s implausible to suppose that a normative fact or a desire could metaphysically create the fact that something is a reason, but more plausible—though still somewhat mysterious—how an act of will—such as that of God—could bring into existence the fact that something is a reason. So one striking interpretation of source voluntarism has it that an act of will is an act of metaphysical creation.²⁰

Exactly how source voluntarism is understood also depends on how the “will” is understood. The “will” in contemporary parlance is usually taken to be a conscious deliberate decision to do something, as when you “steel your will” and do something you don’t want to do, or, as captain, “willingly” go down with your ship. It might be caused by a normative belief about what one has most reason to do, as when one weighs up the pros and cons of two alternatives and finally “wills” to do what one believes is supported by the most reason, or it might be caused by a motivating desire as against such a belief, as in cases of weakness of will.

But the “will” is also sometimes taken to represent the agent himself/herself and “willing” correspondingly taken to be an activity constitutive of agency. You might consciously and deliberately decide to exercise every day, but your will—your agency—is not cooperating. Willing is thus sometimes understood not as a conscious, deliberate decision to do something but as the activity of (rational) agency as such. Perhaps the rational will in this sense—rational agency itself—is what constitutively grounds the fact that something is a reason: a will constrained by rationality is that in virtue of which something is a reason. That was Kant’s view, at least by the lights of some modern interpreters.²¹ Or we might understand willing as the activity of agency involved in putting yourself—your agency—behind something. Perhaps by *willing that something is a reason*—putting your agency behind it as a reason—you can create or construct it as a reason in something like the way God’s willing “Let there be light” creates or constructs the fact that there is light.²² Voluntarists views, while the most intriguing, are among the most mysterious and least understood.

1.3 The Three Questions are Distinct

One common mistake is to move from a view about the content of one’s reasons to a conclusion about the nature of their normativity. So, for example, it might be thought that if reasons are desires then it follows that normativity is a motivating force, for how else can your desire for something be action-guiding except by motivating you to action? Such a view overlooks the logical independence of which considerations are action-guiding from action-guidingness itself. Your desire—or the fact that you want something—can be your reason, but it can be action-guiding by justifying your action, not simply by motivating it. Perhaps wanting the ice cream is normative because it justifies getting some.

Nor would it be correct to move from a view about the content of one’s reasons to the source of their normativity. Some philosophers seem to assume that because reasons are evaluative facts or the facts that subvene them, what grounds the fact that they are reasons must itself be a normative fact. But this assumption must be defended against alternative possibilities. Even if the content of my reason is the fact that doing something would be good or pleasant, such a fact might be normative in virtue of a relation between doing that thing and my desires or dispositions. What makes a fact normative is one thing while the content of that fact is another.²³

It would also be a mistake to move from a view about the source of normativity to a view about the nature of normative force.²⁴ Suppose, for example, that the source of normativity is one’s desires. It would be a mistake to conclude that normativity is thereby a motivational force since it could be

that being properly related to one's desires could be that in virtue of which one's action is justified. Satisfying what you would most want under certain constraints could be what makes your action have the justificatory force of a reason.

The same goes for sliding from source to content. Just because, say, the source of normativity is a realm of irreducibly normative facts, it does not follow that one's reasons are irreducibly normative facts. Perhaps it is an irreducibly normative fact that wanting to do something under certain constraints is, systematically, a reason to do it. Nor does it follow that if desires are what make something a reason, that one's reasons just are the fact that one wants something. The fact that is one's reason should be kept distinct from the fact that this fact is one's reason.²⁵

Finally, views about the nature of normative force are logically independent of views about content or source. It might be thought that reasons must be desires if normativity is a matter of motivation, for what could motivate but desires? But if normativity is a motivating force, it does not follow that reasons are desires, since, as Nagel taught us long ago, desires can themselves be "motivated" by normative beliefs. My belief that cleaning up the mess is good can cause or otherwise necessitate me, as a rational agent, to be motivated to clean it up. My reason is that cleaning up the mess is good, and the normativity it carries could be the motivational force of the desire to clean up the mess that recognition of the goodness of doing so necessitates in a rational being. Of course it is a further substantive question as to whether being rational requires having certain dispositions or desires, such as the desire to do what one believes one has a reason to do, but the point here is that understanding normativity as motivation does not require one to conclude that desires are reasons. Nor does it follow that if normativity is a matter of motivational force that this force must have its source in one's desires or dispositions. It could just be a normative fact that some fact has this motivational force for a rational agent. So the nature of normative force does not tell us where that force comes from.

2 Gridlock

2.1 Three Dominant Types of View about Practical Reasons

While the questions of the content of a practical reason, the nature of normative force, and the source of a reason's normativity are logically independent, there are plausible substantive relationships among their answers. Combinatorially speaking, there are 27 possible groupings of the three leading answers to each question, not all of which make happy bedfellows. Interestingly, philosophers have been overwhelmingly attracted to three particular combinations. I will call them Type 1, Type 2, and Type 3 views.

Type 1 proponents tend to be value-based theorists about content; they think normativity is a *sui generis* justificatory force; and they locate the source of normativity in a realm of normative facts. Suppose you have a reason not to touch the hot poker. Type 1 theorists typically say that the reason is an evaluative fact, such as the fact that touching it will be bad for you, or a fact that subvenes that evaluative fact, such as the fact that touching it will be painful. Sometimes they allow that the fact that one wants something can be what ultimately bears the normativity of a reason, but they tend to think that any such cases are outliers with little philosophical significance.²⁶ They also maintain that the reason not to touch the poker is normative in that it *justifies* and may require not touching it, and this justificatory force comes from a realm of normative facts, such as the fact that being painful is a reason to avoid an experience. Although the questions of content, nature, and source have not always been clearly distinguished, with a little interpretive license proponents of Type 1 views can be said to include Plato (1941), Clarke (1706), Sidgwick (1907), Prichard (1968), Moore (1903), Ross (1930), Nagel (1975), Raz (1986), Scanlon (1998), Dancy (2000), Shafer-Landau (2003), Wallace (2006), Wedgwood (2007), Parfit (2011), and Enoch (2011).

Most Type 2 defenders hold that a systematic study of reasons will reveal that one's reasons are ultimately facts that one wants or would want something under certain formal constraints, but the better versions hold instead that nondesire-based facts are ultimately reasons.²⁷ My wanting to avoid pain or the fact that doing so is painful is my reason not to touch the hot poker, and the normativity of this reason is motivational; reasons are action-guiding in motivating me under certain formal constraints, such as being clear-eyed, calm, fully imaginative, and so on. Moreover, this motivating force has its source in my desires and dispositions, for example, in my desire to have a good life or indeed to avoid painful experiences. It is in virtue of some such desires that the fact that an experience is painful or that I want to avoid it has the motivational power to guide my action. This type of view is arguably held by Hume (1978), Falk (1986), early Foot (1978), Williams (1981), Railton (1989, 2003), Brandt (1996), Darwall (1983), M. Smith (1994), Nichols (2004), M. Schroeder (2007), Tiberius (2008), and probably Rawls (1971).

Finally, Type 3 views are the most explicitly catholic about which sorts of considerations can be one's reasons for action—anything that can be the proper subject of rational willing can be a reason, and there are different versions of what this may include. Your reason to avoid touching the hot poker might be that doing so would be painful or that you don't want pain, and this fact is normative in that it moves your will to action insofar as you are rational. This volitional force is both motivational, because it is what moves the rational agent to action, and justificatory, because it is an expression of the rational will, which is "authoritative" for the agent. Moreover, this force

comes from the will or its activity, constrained by formal requirements of rationality, usually understood as constitutive of rational agency or of action itself. To take one Kantian version, your willing, constitutively constrained by formal requirements of rationality such as consistency and coherence, legislates for itself a principle of action: "Avoid hot poker for the reason that one wants to avoid pain." The willing of this principle then constitutes or creates the fact that wanting to avoid pain is a reason for you to avoid hot poker. Philosophers attracted to views of this type arguably include all divine command theorists as well as Hobbes (1651), Pufendorf (1672), Locke (2003), Kant (1785), and some modern-day neo-Kantian constitutivists such as Korsgaard (1996).

2.2 Why Do These Three Types Persist?

Although the three types of view are not forced upon us, they have dominated the history of theorizing about practical reasons.²⁸ Indeed, the debate about practical reasons can be roughly characterized as one long quarrel among them, with versions of each type taking their turn in the ascendant, only to be temporarily replaced with views of another type, and then coming around again as the favored view.

Of course, the three dominant types may endure because they are the best hypotheses for the truth. But there is another possibility. Perhaps they endure because there is some sense in which their proponents are "talking past one another," and as a result, objections to one type by proponents of another fail to hit their mark. Genuine progress in understanding practical reasons is thus impeded.

One way this might happen is if proponents of each type of are quite literally talking past one another. This is Parfit's diagnosis of the debate in his magisterial *On What Matters*. Parfit, a leading proponent of the first type of view, argues that leading proponents of the second type, such as Bernard Williams, W. D. Falk, and Stephen Darwall, do not even share his *concept* of a practical reason.²⁹ If Parfit is right, what looks like a debate about how to understand practical reasons is not a genuine disagreement at all but rather a pot-pourri of claims about different but related phenomena. Philosophers, despite their sophistication, are having a merely verbal disagreement.

Parfit makes an intriguing and powerful case that this is indeed what is going on. I won't engage with his arguments, which span over hundreds of pages. Instead, I want to suggest a less severe—and I hope less depressing—diagnosis. I believe that proponents of each of the three dominant types of view share the intuitive idea that practical reasons are action-guiding or "count in favor" of action, and that, although they may have different substantive views about what "counting in favor of" amounts to, having this

thin, intuitive notion is sufficient for them to share the concept of a practical reason. If this is right, the question then becomes whether there is nevertheless some explanation as to why, despite this shared concept, theorists often seem to be talking past one another.

2.3 Argumentative Cocoons

We can start with the conjecture that proponents of each type of view seem to take as a starting point within our framework a certain assumption about the nature of normative force, and note that once this assumption is in place, the rest of their view naturally follows. These starting points themselves may derive from prior commitments to naturalism or nonnaturalism about reality.

The Type 1 theorists, for example, seem to anchor their thinking about practical reasons in the substantive assumption that normativity is a *sui generis*, justificatory force. From this assumption, it's quite natural for them to go on to hold that this force has its source in normative facts and that the content of reasons is given by facts about the value or worth of things or the nonevaluative facts that subvene them. From an assumption that normative force is justificatory, a tidy Type 1 package easily follows.

If, instead, normativity is thought to be a motivating force, then it's natural to think that this force has its source in our desires and dispositions, for what could be the source of a motivation other than some other (actual or counterfactual) motivations? And since one is being naturalistic about both the nature of normativity and its source, one might as well be naturalistic about content too and maintain that the ultimate bearers of normativity are natural facts. The genesis of Type 2 views I suspect has this form; defenders start with substantive assumptions about what normative force could plausibly be—it has to be natural—and from there, the rest of their view naturally unfolds.

Finally, it seems plausible that, given larger metaphysical commitments, Type 3 theorists start off wedded to the idea that normative force is essentially something that grips the will. (They have this view because they doubt the capacity of desires and normative facts to explain why anything is authoritative for the agent). Once you think that normativity is a volitional force, then it is natural for you to think that its source is in (rational) willing and that the content of a practical reason can be anything that can be the subject of (rational) willing. Indeed, a Kantian-type package seems almost to fall out of the assumption that normative force is volitional force.

These starting points are not benign. They create an argumentative gridlock among the three types of view because the arguments *for* each view derive their force largely from their respective substantive assumptions about the nature of normative force, while the leading arguments *against* each view

derive their force largely from opposing assumptions about normative force. Thus opponents of one type of view present arguments that presuppose something about the nature of normative force that the strongest arguments for the target view presuppose to be incorrect. In this way, the objections to each type of view fail to hit their mark. By lacking a common substantive view about the nature of normative force, the arguments in the debate fail to engage with what lies at the heart of the disagreement between them—the nature of normative force.

Consider, for instance, what proponents of the first type of view regard as a compelling, and perhaps decisive, objection against views of the second type:

According to Type 2 views, someone with a certain array of desires may have reason to want agony, understood as an intensely disliked sensation, for its own sake. But this is clearly false; no one has reason to want agony for its own sake, whatever his desires.

This is the nub of Parfit's famous "Agony Argument," and it has its roots in the idea that there are goods whose value doesn't depend on what one wants or would want under evaluatively neutral conditions.³⁰ Any view that has as a consequence that someone can in principle have a reason to want agony for its own sake must be rejected. Those sympathetic to the idea that normative force is a *sui generis* justificatory force—that action-guiding force is not itself motivating force—will likely find this argument, especially as Parfit so nicely lays it out, decisive against Type 2 views.

But many proponents of the second type of view find the argument largely untroubling. If the person in question has an intrinsic desire to have agony for its own sake, and if having the desire—or the agony—in no way frustrates or diminishes the satisfaction of any of his other desires or the desires he would have under certain formal constraints, then, yes, indeed, he has a reason to want agony for its own sake. That's exactly as it should be. The agony argument makes no real dent because, for proponents of Type 2 views, normativity—"counting in favor of"—is assumed to be a matter of motivation, and how can someone have a reason—and thus be motivated—not to have a desire if he could not, given certain formal constraints, be motivated not to have such a desire?³¹

Investigation of the argumentative structure of this exchange reveals that Parfit's objection is not really aimed at Type 2 views *tout court* but more specifically at the view of normative force that it assumes to be the case. But since proponents of the second type are, commonly, in the grip of a substantive assumption about what normative force is like, the argument fails to hit its mark. This is not a flaw in the argument but a feature of the argumentative

structure of the debate. By taking a firm substantive stand on the nature of normative force—it must be a motivating force—the intuitive force of objections from the point of view of a different substantive stand—it is a *sui generis* justificatory force—fails to have argumentative power.

Suppose we modify Parfit's objection so that it includes as an explicit premise that normativity is a matter of motivating force—what Parfit's opponents assume to be the case. It would then go something like this:

Suppose, as a substantive hypothesis, that normativity is a matter of motivational force under certain formal constraints. It's possible that someone non-instrumentally wants agony for its own sake under these formal constraints. Since desires are motivating, it follows that this desire could have normative force.³²

So far, this is only an observation about the upshots of a possible psychology under a substantive assumption about normative force, but not yet an objection.

In order to turn the observation into an objection, a further claim is needed, namely, that the assumption that normativity is a matter of motivation is mistaken. This claim might, in turn, be supported by another: "Abstracting from any substantive assumptions about the nature of normative force, a noninstrumental desire for agony for its own sake does not have normative force." This is essentially an appeal to purported intuitions about what could have normative force, apart from any substantive assumptions about what normative force might be like. But now we can see that the Parfitian objection is not against the second type of view about practical reasons *tout court* but more particularly against its answer to the question of the nature of normative force. We might reformulate the nub of Parfit's Agony Argument like this:

Type 2 views make a mistake when they assume that normativity is a motivating force because intuitively, according to the thin, intuitive idea of "counting in favor of," a non-instrumental desire for agony for its own sake does not have normative force, which it could have if normativity were a motivating force.

With the target of Parfit's argument more clearly in view, however, we can see why there is an impasse. Proponents of the second type claim not to have the intuition that it is odd to say that the noninstrumental desire for agony for its own sake is not "normative" in the neutral sense, that is, that it doesn't count in favor in the thin, intuitive sense upon which all interlocutors to the debate can agree. As they see things, the best account of normativity in the neutral

sense is as a motivating force. This is not to say they *define* normativity in these terms; rather their substantive views about normativity make it hard for them to feel the intuitive pull of Parfit's example.

In short, for Parfit's objection to hit its target, it must be supposed that normativity is not simply a matter of what one would be motivated to do under suitable formal constraints. This is because the force of the objection is supposed to be that we intuitively think that someone does not have a reason even though he/she is motivated under these constraints. But the force of the objection presupposes a substantive view of the nature of normativity that is not shared by the view against which it is leveled.

Having belabored discussion of this first example, we can move more quickly through others since the core problem is the same. Consider the leading objection that proponents of Type 3 views have made against proponents of Type 1 views:

According to Type 1 views, someone can have a reason to do something even though this reason in no way engages his will. But this is absurd; reasons have to "get a grip" on those for whom they are reasons in order for them to be reasons for that person.

This is more or less Korsgaard's famous argument against normative realists, and it has its roots in the idea that nothing can be a reason for someone unless it can engage his rational agency, where "rational" here means not substantive rationality but what Scanlon has called "structural" rationality, the rationality of requirements.³³ Those sympathetic to the idea that normative force is a force that moves the will, tend to find this argument, properly spelled out, compelling.

But proponents of Type 1 views find the enthusiasm with which the objection is lodged against them somewhat baffling. Why should one think that a *sui generis* justificatory force must engage the will of the agent? Insofar as reasons get a grip on agents, they do so when such agents are rational—in the substantive sense, that is, recognizing and responding to those reasons. Asking why something's being a reason should have a grip on a rational agent hardly makes sense. What it is to be rational in this sense is to be gripped by one's reasons. The objection seems to presuppose that normativity must be some kind of volitional or motivational force, but since it isn't, the argument misses its mark.

The same goes for other familiar arguments against the first type of view: the metaphysical and epistemological worries that Mackie made familiar. Those arguments get their punch from the assumption that normativity must be a natural force—what else, could it be and how else could we come to know what has it? If, however, we remain neutral on the nature of normative

force, resting content with the thought that, natural or not, normativity is not more queer than other sorts of necessity we can't do without, we pull their punch.

Finally, consider the leading objection made by proponents of the first and second types of view to a version of the third:

According to certain versions of Type 3 views, having a reason to *x* is nothing more than willing a principle of *x*-ing in accord with the requirements of practical rationality. But the question, arises, "Why follow the requirements of practical rationality?" And the answer given by proponents of the third view has to be either, "it's good to follow the requirements of practical reason," in which case the third view appeals to something with normativity that goes being willing a principle in accord with the requirements of practical reason—the "goodness" of following such requirements—, or "there's a requirement of practical reason that one follow the requirements of practical reason," in which case the question arises all over again, and we are led to an infinite regress of requirements of practical reason. The third view cannot successfully answer the question, "What reason do we have to will in accordance with the requirements of practical rationality?"

This "regress argument," among other objections, has led the bulk of philosophers, most of whom endorse some version of Type 1 or Type 2 views, to dismiss the third as a nonstarter.³⁴ But to proponents of Type 3 views, this objection is simply misguided. If normativity is a *sui generis* volitional force deriving from the rational will, then there can be constitutive requirements governing the proper exercise of the will which can make something a reason. By hypothesis, something doesn't count as a rational willing or, indeed, as nondefective action at all, unless it conforms to the requirements of practical reason; that's part of what it is to be rational willing or a nondefective action. Thus, since having a reason just is willing or nondefectively acting in accord with rational requirements, asking what reason one has to will in accord with rational requirements is a nonsensical question. Once one takes as given that normativity is just volitional force—and more specifically—that something's having this volitional force is a matter of one's willing according to the rational law, the regress objection doesn't seem to get started.³⁵

Our purpose is not to engage in substantive debate about the three dominant types of view but only to give some examples of how the argumentative force of some of the leading objections to each are hostage to assumptions about the nature of normative force, assumptions not shared by their target views. We have seen how, because of unshared assumptions about normative force, these objections fail to hit their mark.

2.4 Argumentative Gridlock

Much of philosophical debate proceeds as follows. Everyone shares the concept—for example, “knowledge”—at issue. Views about knowledge, then, are about a common matter of inquiry. Proponents of theory A of the concept will offer a strong criticism of theory B. Proponents of theory B will do their best to undermine or answer this criticism, sometimes by clarifying, deepening, or modifying their view, and will endeavor to show that their theory, despite its difficulties, has greater theoretical merit overall than that of their rivals. Proponents of theory A will do the same. There is no argumentative gridlock but genuine back-and-forth engagement over issues with resulting modifications to existing views.

The debate about practical reasons has a different structure. By hypothesis, philosophers share the concept of a normative practical reason—it is a consideration that “counts in favor” of performing a certain action or having a certain attitude. In fashioning theories of practical reasons, however, they take a substantive position on the nature of normative force and from this the rest of their view—on content and source—naturally unfolds. But as we have seen, this move from an assumption about the nature of normative force to a full view about practical reasons is problematic in that it creates an argumentative cocoon around the view to which it naturally gives rise. Proponents of one type cannot effectively criticize opponents of another since their criticisms presuppose a view about the nature of normative force rejected by their opponents. This is so even though they share the concept of a practical reason as something that “counts in favor” of an action or attitude.

Now one way out of this gridlock would be to recognize that what is essentially at issue in the debate between the three types of view is the nature of normative force. But there’s the rub. We have no idea how to make progress on the question of the nature of normative force. The nature of normative force is plausibly one of those “hardest” problems, like the nature of substance or consciousness, about which it may be best to make certain assumptions in an attempt to move forward in understanding other matters. It is thus unsurprising that philosophers have helped themselves to assumptions about the nature of “counting in favor of” in developing theories of practical reasons. Taking the plunge on a hard problem in order to see where it leads is often a good strategy for tackling a philosophical problem. But, as I have suggested in the case of understanding practical reasons, substantive assumptions about normative force are far from benign; they have had the effect of imposing a structure on the debate that leads to argumentative gridlock.

This argumentative situation bears some similarity to an impasse in the debate about abortion. When a pro-choicer appeals to the absolute right, *ceteris paribus*, of a woman to determine what happens to her own body in arguing against prohibitions to second-trimester abortion in cases of rape, a pro-lifer

might fail to see that this reason has absolute strength not because he defines or conceptually fiat a second-trimester fetus as a person but because he has a substantive metaphysical view about what such a fetus is that precludes him from seeing how there could be an absolute right of another to destroy it. In both cases, we have substantive metaphysical views standing in the way of appreciating what are supposed to be claims that favor one substantive view over another. And in both cases we are genuinely stymied over how to make progress on that metaphysical question.

As is common in many philosophical debates, deep metaphysical questions can infect other debates in which they strictly have no proper place. For example, sometimes first-order normative theorizing is infected with assumptions about whether everything is natural or whether there are nonnatural facts or properties. Theories about what makes something morally good, for instance, should not be driven by assumptions about whether everything is natural—so that whatever makes things morally good must be natural. Instead, normative theorizing should be neutral on this deep metaphysical question and instead be attuned to the subtleties of higher level facts that might be relevant to determining what makes something morally good as a substantive, normative matter. Something similar may be at work in the present debate. Deep metaphysical assumptions about the nature of what there could be—everything is at bottom natural, for instance—may underwrite substantive assumptions about the nature of normativity that in turn wreak havoc in debate about the metaphysics of practical reasons.

3 Coda: A Focus on Source?

If the state of inquiry into practical reasons is as I have described it, the question then arises, What now? We can again turn to our framework for a possible answer.

If my diagnosis of the gridlock is correct, and if I'm right that we are presently ill-equipped to make progress on the nature of normative force, we might try a different tack. To break the stranglehold of the three dominant types of view, we might abstract away from assumptions about the nature of normative force—and, more generally, stop allowing the metaphysical question of what there ultimately could be infect higher level debates—and focus instead on the other questions of our framework, the questions of the content of reasons and of their normative source. By focusing on these questions without already having in hand a substantive stand on the nature of normative force, we will naturally bring to bear different sorts of considerations in defending possible views. So, for instance, we might focus our attention on the nature of desires to see how they could be systematic bearers or grounds

of normativity.³⁶ Or we might try to understand the will. The last extended book—length treatment of the will dates from over 30 years ago.³⁷

I suggest that we focus on the question of source. The strategy of abstracting away from assumptions about normative force and focusing on normative source has not gained much traction so far, I believe, because the question of source is thought to be obscure. This chapter is one attempt to try to make it less so. The question, “What makes a consideration ‘count in favor’ in the thin, intuitive sense we can all grasp?” is not one that need stymie us.

Whether thinking about normative source will help illuminate practical reasons remains to be seen. I want to end by describing one way in which a focus on source might naturally suggest a new and fruitful way of thinking about practical reasons.

Philosophers tend to assume that the way to individuate a reason is the same way to individuate a cause—by its “content.” Since causality is univocal in its source, this is a good strategy for cause. But if we are to be neutral as the nature of normativity, we should be open to the possibility that there are multiple sources of normativity, a possibility so far (to my knowledge) ignored in the debate about practical reasons.³⁸ If there can be multiple sources of normativity, then a plausible way to individuate reasons will be by their content *and* source. The same “content”—a single fact or consideration—can be the content of two quite different reasons if what *makes* that consideration action-guiding is different in each case. The same point can be put less dramatically by emphasizing that there are two different ways in which something can be a reason—via normative fact and via one’s desires. We need not talk about having two different reasons with the same content but rather a reason with a specific content, which may be a reason in one of two ways.

Take, for example, the fact that an experience is painful. This fact may be action-guiding in virtue of the normative fact that being painful is a reason to avoid things with that feature. What makes the fact of being painful normative is some normative fact. Thus this fact—with this normative source—is a reason not to touch the hot poker. But perhaps the fact that an experience is painful can also be action-guiding in virtue of one’s intrinsic desire to avoid pain. Your intrinsic desire to avoid pain might make the fact that it’s painful action-guiding. In this way, the very same content—that it’s painful—can be two different practical reasons.³⁹ Or, again, more concessively, being painful can be a reason in one of two ways.

Notice that this result—that those who have an intrinsic desire to avoid pain for its own sake have more reason to avoid it than those lacking such a desire—accounts for our intuition that, even if everyone has a reason to avoid pain for its own sake, regardless of her desires to have or not to have such disliked sensations, someone with an intrinsic desire to avoid pain for its own sake would seem to have *more* reason than someone without such a desire.⁴⁰

Understanding reasons by direct appeal to their normative source gives us a nice way of accounting for this difference.

A focus on source not only suggests a way of understanding reasons in terms of source and the possibility of multiple normative sources, but also naturally occasions inquiry into how, if reasons can have different normative sources, how those sources are related. Perhaps when reasons deriving from one source “run out”—when they fail determinately to answer the question of what one has most reason to do—reasons deriving from another source can step in to pick up the slack. This is a view I favor but there are many other intriguing possibilities.

Finally, thinking in a sustained way about normative source pushes us to consider more closely the role of agency in understanding practical reasons. This is because asking where normativity comes from highlights the possibility that agency is the ground of action-guidingness. This would mean bringing the philosophy of action and agency back into the center of debate about practical reasons which, many would argue, is where it belongs⁴¹

Notes

- 1 There is a third, traditional branch of ethics, applied or ‘practical’ ethics, investigation of what to do in concrete, particular ethical situations. In the past few decades, this branch of ethics has shown increasing sensitivity to other areas of philosophy. It too may develop in new ways by being influenced by the philosophy of practical reason.
- 2 This describes, I think, the situation in the debate between Derek Parfit and many of his critics in a forthcoming collection edited by Peter Singer about Parfit’s meta-ethical claims and arguments. See Singer (forthcoming).
- 3 It is a further substantive question whether these meta-normative questions reduce to normative ones as, for example, quietists think. But we need not get into that question here.
- 4 Williams 1981; but compare Dancy 2000, 2004 who argues that even motivating reasons, if they are to be good normative reason for action, must be value-based facts.
- 5 For example, Parfit 2011; Scanlon 1998; Raz 1997.
- 6 Chang 2004. This was also arguably the view in Parfit 1984, but repudiated in Parfit 2011.
- 7 Williams 1981 et al. An interesting development of Williams’ view understands the concept of a reason to be explained in terms of the idea of *advice*. See Manne (forthcoming).
- 8 Scanlon 1998; Raz 1999; Dancy 2000; Shafer-Landau 2003; Enoch 2011; Parfit 2011.
- 9 Judith Thomson thinks that something is practical normative insofar as it “lends [some kind of] weight” to the proposition that one should do what the reason is supposedly a reason for. (Thomson 2008, 156). Others make explicit that a relation of evidential support for the proposition that one should do something is what makes a consideration a reason, thus embracing a unified view of the nature of normativity for both practical and theoretical reasons (Kearns and Star 2009).

- 10 Korsgaard arguably understands normativity in this way—either as a conjunction of justificatory and motivating force or as a *sui generis* volitional force with both justificatory and motivational aspects. See Korsgaard 1996, 2008.
- 11 See, for example, Korsgaard 1996, 2008.
- 12 The failure to distinguish the source question from the “conditions” questions has led some philosophers to suggest that “a reason” is *anything* that is relevant to the explanation of why someone should do something. This is fine as far as stipulating the use of a term goes, but it has a downside; it obscures the different explanatory roles different facts relevant to a complete explanation of why someone should do something can play, the roles that I here have marked out as “content,” nature, and source. An analogous distinction in this respect—along with analogous controversy—arises between the cause “proper” and the conditions under which something is a cause. See, for example, Mackie 1980 vs Davidson 1963.
- 13 See also my 2013a from which this discussion is drawn.
- 14 The idea of “ground” was introduced into contemporary metaphysics by Fine 2001 (see also his 1991 and 1994 where he argues that some metaphysical explanations are not modal).
- 15 Metaphysical creation may, in the end, reduce to some form of constitution. I give a fuller defense of the idea in my 2013a, arguing how it is not an instance of causation in particular. More needs to be said about each of these grounding relations, especially the last, but my aim is only to outline the ones I believe are relevant for understanding practical normativity.
- 16 Of course, it is perfectly consistent with source externalism position to hold what makes the pain a bad experience is something subjective, namely that it is an experience one intensely dislikes. Some philosophers seem to assume that if a reason involves subjective states, like enjoyment, those reasons are internalist in source (e.g. Schroeder 2008), but this does not follow. See also Parfit 2011.
- 17 Some source externalists are pluralists, holding that sometimes the fact that something is a reason is self-grounded and sometimes it is constituted by an evaluative fact. See, for example, Raz 1999. What sort of externalist view one holds depends on one’s views about the logical priority between values and reasons. If both are primitive normative phenomena one is likely to be a pluralist source externalist.
- 18 This appears to be Michael Smith’s view. See Smith forthcoming.
- 19 See Schneewind 1998. My invocation of Kant here is admittedly controversial since on most readings of Kant, willing is not something genuinely “up to us.” See also Korsgaard 1996, 2008.
- 20 Those who are attracted to source voluntarism but are skeptical of Korsgaard’s claim that guidance by the Categorical Imperative is constitutive of action itself might find a defense of source voluntarism that takes the grounding relation to be one of metaphysical creation, not constitution. I try to offer the beginnings of such a defense in various articles cited below.
- 21 See Korsgaard 1996 and 2008 for a development of this view.
- 22 I defend such a view in Chang 2009, 2013a, 2013b, Ms.
- 23 For instance, Parfit 2011 and Scanlon 1998 think that what makes the fact that the medicine is effective a reason to take it is a normative fact that such a fact is a reason to take it, but Railton and Williams 1981 think what makes that fact a reason is some relation between taking the medicine and one’s desires. This difference between Parfit and Scanlon, on the one hand, and Railton and Williams, on the other, is sharpened by understanding their disagreement as one about the source of normativity, not about which facts can be reasons, which is how their disagreement is often presented.
- 24 See also the “externalist subjective theory” in Parfit 2011.

- 25 This is one of the main points I attempt to make in Chang 2004.
- 26 See for example Scanlon 1998, ch. 1; Parfit 2011, 52–6; Raz 1997. These exceptions are discussed in Chang 2004.
- 27 Most Type 2 theorists elide the question of what makes some fact a reason with the question of what is the reason, but once this distinction is made, perhaps most would allow, as Type 1 theorists do, that nondesire-based facts can or must be ultimate reasons.
- 28 This is not to say that every view about practical reasons is a version of one of the three dominant views. Some neo-Kantians, for example, have views that cut across the dominant views, and are, in particular, externalists about source. Thomas Hill 1991, 2002 thinks that there is a normative fact—being rational is acting on such-and-such consideration—that is that in virtue of which the consideration is a reason. Other neo-Kantians are externalist about source in other ways, grounding normativity in the intrinsic value of the good will (Herman 1993) or of the person herself (Anderson 1993). Julia Markovitz has recently suggested a view according to which the source of normativity is in our desires (or some relation to them) but normativity is a *sui generis* justificatory force (Markovitz forthcoming). Other theories not falling neatly into any of the types of view include expressivism, which is primarily a theory of normative judgments, and error theories, which might be understood as “negative” theories—for instance, as the negation of some of the claims of Type 1 views.
- 29 Parfit 2011.
- 30 This is Parfit’s ‘second’ agony argument, but since I believe it is stronger than the ‘first’, I discuss only this one.
- 31 I believe this is the central thought behind Williams’ declaration that he cannot understand how there could be ‘external’ reasons, reasons to do things that one could not, through sound deliberation, come to be motivated to do (Williams 1981).
- 32 The requirement that the desire be noninstrumental blocks evil demon cases in which the evil demon will torture you unless you desire agony for its own sake.
- 33 Scanlon 1998; Broome 2004.
- 34 See Railton 2004, Scanlon 2003, and relatedly Enoch 2006.
- 35 See especially Korsgaard 2009 who argues that since action is inescapable, and it must as a constitutive matter be guided by the rational law, it makes no sense to ask what reason we have to follow the rational law. But this constitutivist maneuver does not, it seems to me, succeed. It can be constitutive of torture that pain be applied in an excruciating way, and there can be creatures that for whom torture is nonoptional. This does not block the question, What reason is there to apply pain in an excruciating way? I discuss this objection further in my 2013a.
- 36 I interpret recent work by Peter Railton as developing this line of thought. Type 2 theorists, it seems to me, could make the most progress by taking this Railtonian approach. The crucial issue then becomes whether Type 2 theorists can get us the robust kind of normativity that Type 1 theorists (and common sense) suggests there is.
- 37 I am thinking of O’Shaughnessy’s two-volume set from 1980.
- 38 I seem to be the only philosopher, to my knowledge, who has taken this possibility seriously. It would be nice to have some company. See my 2009, 2013a, 2013b, Ms.
- 39 An imperfect but suggestive analogy from physics: a single object, like a magnet, can exert more than one kind of force on an object—it can be an electromagnetic force and it can be a gravitational one. In a roughly analogous way, a single fact can be two kinds of reason, where the “kind” is not marked by the kind of normative force but rather by that in virtue of which the fact has normative force.

- 40 Of course, value-based opponents would say that this fact is explained by the fact that for those who have the desire to avoid pain the experience would be worse, and it's being worse (or the ways in which it is worse) is the reason. I try to counter such an argument in my 2004 work.
- 41 Thanks to Kit Fine and Derek Parfit for helpful comments and to the editors of this volume for helpful editorial advice.

Bibliography

- Anderson, E., 1993. *Value in Ethics and Economics*. Cambridge, MA: Harvard University Press.
- Brandt, R., 1996. *Facts, Values, and Morality*. Cambridge: CUP.
- Broome, J., 2004. "Reasons." In R. J. Wallace, P. Pettit, S. Scheffler, and M. Smith, eds, 2004.
- Chang, R., Ms. "Do We Have Normative Powers?" *Draft*.
- , 2013b. "Commitment, Reasons, and the Will." *Oxford Studies in Metaethics*, 8, pp. 74–113.
- , 2013a. "Grounding Practical Normativity: Going Hybrid." *Philosophical Studies*, 164 (1), pp. 163–87.
- , 2009. "Voluntarist Reasons and the Sources of Normativity." In D. Sobel and S. Wall, eds, *Practical Reason and Action*. Cambridge: CUP, pp. 243–71.
- , 2004. "Can Desires Provide Reasons for Action?" In R. J. Wallace, P. Pettit, S. Scheffler, and M. Smith, eds, *Reasons and Values: Themes from Joseph Raz*. Oxford: OUP.
- Clarke, S., 1706. *Discourse Concerning the Unchangeable Obligations of Natural Religion, and the Truth and Certainty of the Christian Revelation*. In D. D. Raphael, ed., 1991, pp. 191–225.
- Dancy, J., 2004. *Ethics without Principles*. Oxford: Clarendon.
- , 2000. *Practical Reality*. Oxford: OUP.
- Darwall, S., 1983. *Impartial Reason*. Ithaca: Cornell University Press.
- Davidson, D., 1963. "Actions, Reasons, Causes." *Essays on Actions and Events*. Oxford: Clarendon.
- Enoch, D., 2011. *Taking Morality Seriously: A Defense of Robust Realism*. Oxford: OUP.
- , 2006. "Agency, Shmagency: Why Normativity Won't Come from What is Constitutive of Agency." *Philosophical Review*, 115, pp. 169–98.
- Falk, W. D., 1986. *Ought, Reasons, and Morality: The Collected Papers of W. D. Falk*. Ed. W. D. Falk and K. Baier. Ithaca: Cornell University Press.
- Fine, K., 2001. "The Question of Realism." *Philosophers' Imprint*, 1 (1), pp. 1–30.
- , 1994. "Essence and Modality." *Philosophical Perspectives*, 8, pp. 1–16.
- , 1991. "The Study of Ontology." *Nous*, 25 (3), pp. 263–94.
- Foot, P., 1978. "Morality as a System of Hypothetical Imperatives." Reprinted in her *Virtues and Vices*. Oxford: Basil Blackwell.
- Herman, B., 1993. *The Practice of Moral Judgment*. Cambridge, MA: Harvard University Press.
- Hill, T., 2002. *Human Welfare and Moral Worth: Kantian Perspectives*. Oxford: OUP.
- , 1991. *Autonomy and Self-Respect*. Cambridge: CUP.

- Hobbes, T., 1651/1988. *Leviathan*. New York: Prometheus Books.
- Huemer, M., 2005. *Ethical Intuitionism*. New York: Palgrave MacMillan.
- Hume, D., 1739/1978. *A Treatise of Human Nature*. Edited by L. A. Selby-Bigge and P. H. Nidditch. Oxford: OUP.
- Kant, I. 1785/1959. *Foundations of the Metaphysics of Morals*. Translated by Beck, Indianapolis: Bobbs-Merrill C.
- Kearns, S. and Star, D., 2009. "Reasons as Evidence." *Oxford Studies in Metaethics*, 4, pp. 215–42.
- Korsgaard, C., 2009. *Self-Constitution: Agency, Identity, and Integrity*. Oxford: OUP.
- , 2008. *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*. Oxford: OUP.
- , 1996. *The Sources of Normativity*. Cambridge: CUP.
- Locke, J., 1689/2003. *Two Treatises of Government, and, A Letter Concerning Toleration*. Edited by I. Shapiro. London: Yale University Press.
- Mackie, J., 1980. *The Cement of the Universe*. Oxford: OUP.
- , 1977. *Ethics: Inventing Right and Wrong*. Harmondsworth: Penguin Books.
- Manne, K., Forthcoming. "Internalism about Reasons: Sad but True?" *Philosophical Studies*.
- Markovits, J., Forthcoming. *Moral Reasons*. Oxford: OUP.
- Moore, G. E., 1971. *Principia Ethica*. Cambridge: CUP.
- Nagel, T., 1975. *The Possibility of Altruism*. Oxford: Clarendon.
- Nichols, S., 2004. *Sentimental Rules*. Oxford: OUP.
- Nozick, R., 1981. *Philosophical Explanations*. Cambridge, MA: Harvard University Press.
- Parfit, D., 2011. *On What Matters*. Oxford: OUP.
- , 1986. *Reasons and Persons*. Oxford: OUP.
- Plato, 1941. *The Republic*. Translated by B. Jowett. New York: The Modern Library.
- Prichard, H. A., 1968. *Moral Obligation and Duty and Interest: Essays and Lectures by H.A. Prichard*. Edited by W. D. Ross and J. O. Urmson. Oxford: OUP.
- Pufendorf, S. von, 1672. *On the Law of Nature and of Nations* as discussed in Schneewind 1998, pp. 88–9.
- Railton, P., 2004. "How to Engage Reason: The Problem of Regress." In R. J. Wallace, P. Pettit, S. Scheffler, and M. Smith, eds 2004.
- , 2003. *Facts, Values, and Norms: Essays Toward a Morality of Consequence*. Cambridge: CUP, 2003.
- , 1989. "Naturalism and Prescriptivity." *Social Philosophy and Policy*, 7, pp. 151–71.
- Rawls, J., 1971. *A Theory of Justice*. Cambridge, MA: Belknap Press.
- Raz, J., 1999. *Engaging Reason: On the Theory of Value and Action*. Oxford: OUP.
- , 1997. "Incommensurability and Agency." In R. Chang, ed., 1997. Reprinted in Raz 1999, ch. 3.
- , 1986. *The Morality of Freedom*. Oxford: OUP.
- Ross, W. D., 1930. *The Right and the Good*. Oxford: Clarendon.
- Scanlon, T., 2003. "Metaphysics and Morals." *Proceedings and Addresses of the American Philosophical Association*, 77, pp. 7–22.
- , 1998. *What We Owe to Each Other*. Cambridge, MA: Belknap Press.
- Schroeder, M., 2008. *Slaves of the Passions*. Oxford: OUP.
- Schneewind, J., 1998. *The Invention of Autonomy*. New York: CUP.

- Shafer-Landau, R., 2003. *Moral Realism: A Defense*. Oxford: Clarendon.
- Sidgwick, H., 1907. *The Methods of Ethics*. 7th edn (1st edn 1874, reprinted 1981). Indianapolis: Hackett Publishing.
- Singer, P., ed., Forthcoming. *Does Anything Really Matter? Parfit on Objectivity*. Oxford: OUP.
- Smith, M., 2013. "Parfit's Mistaken Meta-ethics." In Singer, ed., forthcoming.
- , 1994. *The Moral Problem*. Oxford: Blackwell Publishing.
- Tiberius, V., 2008. *The Reflective Life: Living Wisely with Our Limits*. Oxford: OUP.
- Thomson, J. J., 2008. *Normativity*. Chicago: Open Court.
- Trogon, K., Forthcoming. "An Introduction to Grounding." In M. Hoeltje, B. Schnieder, and A. Steinberg, eds, *Dependence*. Munich: Philosophia Verlag.
- Velleman, D., 2000. *The Possibility of Practical Reason*. Oxford: Clarendon.
- Wallace, R. J., 2006. *Normativity and the Will*. Oxford: OUP.
- Wallace, R. J., Pettit, P., Scheffler, S., and Smith, M., eds, 2004. *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*. Oxford: OUP.
- Wedgwood, R., 2007. *The Nature of Normativity*. Oxford: OUP.
- Williams, B., 1995. "Internal Reasons and the Obscurity of Blame." *Making Sense of Humanity and Other Philosophical Papers, 1982–1993*. Cambridge: CUP.
- , 1981. "Internal and External Reasons." *Moral Luck*. Cambridge: CUP.

29 Moral Demands and Ethical Theory: The Case of Consequentialism

Attila Tanyi

1 Introduction

Morality is demanding; this is a platitude. It is thus no surprise when we find that moral theories too, when we look into what they require, turn out to be demanding. However, there is one leading moral theory – consequentialism – which is confronted to by a demandingness *problem*. The demands the theory makes on us are so great (or so the argument runs) that consequentialism must be rejected. It simply requires *too much* of us.

This objection gives rise to a number of pressing questions. Is it right to claim that consequentialism makes excessive demands? Is there a limit on how demanding morality can be? Why single out consequentialism? Is it the only moral theory that makes unacceptably high demands? If the demandingness problem is real, what is the consequentialist's best response? This chapter sets out to answer these questions (or at least point to how they could be answered).

The chapter has the following structure. I will first present the objection (section II), then explain how it differs from other objections (section III) and why it targets consequentialism in particular (section IV). We will see that it is not at all easy to explain why the trouble with consequentialism is on account of its demandingness *only*. After this, I will present the objection in a more formal way that helps me to introduce the different ways of responding to it (section V). In the remainder of the chapter, I will discuss three (relatively) underexplored responses: (1) a response that introduces a new, multidimensional version of consequentialism (section VI); (2) a response that focuses on the role of institutions in lowering demands on individuals by introducing an institutional division of labor (section VII); and (3) response that changes the focus to reasons and their connection to consequentialist demands (section VIII). I then end with a brief summary and make some concluding remarks (section IX).

2 Consequentialism and Demandingness

Start with the targeted theory.¹ Consequentialism, in its most general sense, is the view that normative properties depend only on consequences. This general approach can be applied at different levels to different normative properties of different kinds of things, but the most prominent example is consequentialism about the *moral rightness* of acts. This (moral) consequentialism holds that whether an act is morally right depends only on the consequences of that act or of something related to that act, such as the motive behind the act or a general rule requiring acts of the same kind, as judged from an impersonal perspective.

The paradigm case of moral consequentialism is utilitarianism, whose classic proponents were Jeremy Bentham, John Stuart Mill, and Henry Sidgwick. These classical utilitarians were all *act-consequentialist*: They held that whether an act is morally right or wrong depends only on its consequences (as opposed to the circumstances or the intrinsic nature of the act or anything that happens before the act or anything that relates to the act). They were *utilitarians* because they advocated consequentialism with a welfarist theory of value, that is, a theory that focuses on human welfare, well-being, or happiness as the relevant consequence. And since they understood happiness in terms of the balance of the amount of pleasure over pain, they were also *hedonists*. The demandingness objection has originally targeted these classical utilitarians, but can be employed against any form of act-consequentialism.²

What exactly does the objection say? Discussions of the objection normally begin with short stories like the following two:³

The Envelope. On your desk is an envelope addressed to a reputable charity seeking donations to save the lives of victims of a famine or other natural disaster. *Utilitarianism* says you should give *all* your money to this charity, as each dollar will produce more happiness in their hands than you could possibly produce by spending it in any other way.

Your Money and Charity. You are wondering whether to spend a pound on chocolate for yourself or to give it to a certain charity. You know that this charity is unusually effective and that even a small contribution can help them save a child from some crippling and painful illness. Since you obviously do more good by saving a child from illness than by eating a piece of chocolate, you ought to give the pound to charity. However, if you repeat this utilitarian reasoning every time you have a pound to spare, you will end up very poor indeed.

The first story, *The Envelope*, gives us the traditional version of the objection with one large consequentialist demand countering whatever else the agent

might be planning to do. The second story, *Your Money and Charity*, is designed to show how small but iterated demands can add up to altogether excessive demands.⁴ In either the case, the message is the same: the demands of consequentialism are excessive and, therefore, objectionable.⁵

Let us now have a closer look at the structure of the objection. It is built upon two pillars: (1), that consequentialism is excessively demanding, and (2) that an adequate morality cannot be excessively demanding. Consequentialism requires the agent to promote the good (consequences) until the point where further efforts would burden the agent as much as they would benefit others. However, the situation that determines what would be best overall is far from ideal: today's world involves, for example, significant levels of poverty that prevailing levels of charitable donations are insufficient to eradicate.⁶ Given that acting to alleviate poverty is likely to have, in sum, more positive consequences than pursuing individual goals and projects, it seems unavoidable that, if one fully accepts consequentialism, one must devote most of one's resources to humanitarian work. Both *The Envelope* and *Your Money and Charity* make the same point in their own, more particular way. At the same time, so the objection assumes, most people have a firmly held judgment that this cannot be right, that people should not be required to sacrifice their lives for morality. This is the second pillar of the objection. Its function is to ground a constraint on admissible moral theories requiring them to avoid excessive demands. If they do not, the conclusion follows that these theories should not be allowed to guide people's conduct.

3 Demandingness and Other Objections

The demandingness objection makes a simple case against consequentialism: Since the consequentialist agent is required to maximize the overall balance of good consequences, this is excessively demanding (in our present circumstances), hence objectionable (because one should be given the opportunity to have a life outside morality). To get a clearer grasp of the objection, it is useful to contrast it with other objections to consequentialism that are often bundled together with it. This will ultimately also contribute to our understanding of why consequentialism is singled out by the objection as its sole target.

Here is a famous case from Bernard Williams (1973a: 98–9):

Jim and the Indians. Jim finds himself in the central square of a small South American town. Tied up against the wall are a row of twenty Indians, most terrified, a few defiant, in front of them several armed men in uniform. A heavy man in a sweat-stained khaki shirt turns out to be the captain in charge and, after a good deal of questioning of Jim which establishes that he got there by accident while on a botanical expedition,

explains that the Indians are a random group of the inhabitants who, after recent acts of protest against the government, are just about to be killed to remind other possible protestors of the advantages of not protesting. However, since Jim is an honoured visitor from another land, the captain is happy to offer him a guest's privilege of killing one of the Indians himself. If Jim accepts, then as a special mark of the occasion, the other Indians will be let off. Of course, if Jim refuses, then there is no special occasion, and Pedro here will do what he was about to do when Jim arrived, and kill them all. Jim, with some desperate recollection of schoolboy fiction, wonders whether if he got hold of a gun, he could hold the captain, Pedro and the rest of the soldiers to threat, but it is quite clear from the set-up that nothing of that kind is going to work: any attempt at that sort of thing will mean that all the Indians will be killed, and himself. The men against the wall, and the other villagers, understand the situation, and are obviously begging him to accept. What should he do?

In sum, Jim gets to choose between two actions: (i) not killing anyone himself, yet thereby causing the death of twenty villagers. Or (ii) killing one of the villagers himself, thereby ensuring that the others go free. Williams argues that a utilitarian cannot avoid that conclusion that (i) is the morally required course of action: Jim should kill one villager.

However, this conclusion can be found troubling for three reasons; Williams himself mentions two. One, the utilitarian reasoning makes it clear that it does not matter how certain consequences are produced: Whether Jim kills one villager, or Pedro, as a result of Jim's refusal, kills all twenty, matters only to the extent that in the latter case the consequences are worse. Utilitarianism is insensitive to the distinction between *doing* and *allowing*, which in this particular case translates into the distinction between *killing* someone and *allowing* them to die (by failing to prevent Pedro's act). The result is what Williams (1973a: 95) calls *negative responsibility*: "that if I am ever responsible for anything, then I must be just as responsible for things that I allow or fail to prevent, as I am for things that I myself, in the more everyday restricted sense, bring about."

However, Williams argues that the fact that utilitarians cannot escape endorsing negative responsibility is a flaw of the utilitarian theory. According to Williams, it matters morally whether you cause an outcome actively (i.e., by producing a certain course of actions) or passively (by refraining to act in a certain way and thus allowing it to happen). It also matters if the chain of causal events that produces an outcome contains someone else's act and decision or not, that is, if an outcome includes someone else's doing or solely my own doing. In short, "each of us is specially responsible for what *he* does, rather than for what other people do" (ibid.: 99). Utilitarianism does not respect this important moral insight and therefore it cannot be the correct moral theory.

So this is one problem with consequentialism. But Williams does not stop here. He argues that endorsing negative responsibility leads to *alienation* from one's own life projects and, ultimately, to the *disintegration* of the self (ibid.: 116–17):⁷

The point is that [the agent] is identified with his actions as flowing from projects or attitudes which . . . he takes seriously at the deepest level, as what his life is about . . . It is absurd to demand of such a man, when the sums come in from the utility network which the projects of others have in part determined, that he should just step aside from his own project and decision and acknowledge the decision which utilitarian calculation requires. It is to alienate him in a real sense from his actions and the source of his action in his own convictions. It is to make him into a channel between the input of everyone's projects, including his own, and an output of optimific decision; but this is to neglect the extent to which *his* actions and *his* decisions have to be seen as the actions and decisions which flow from the projects and attitudes with which he is most closely identified. It is thus, in the most literal sense, an attack on his integrity.

Negative responsibility arises from consequentialism's commitment that only consequences matter, not how and by whom they are produced: what count are the (valuable) states of affairs produced and nothing else. But this means that a utilitarian who is committed to thinking in this way is also committed to look at his or her own projects in the same way: not as in any particular sense *his or her own*, but as only one among many others that matter only to the extent that when satisfied, pursued, accomplished, and so on, they produce valuable states of affairs. However, Williams (ibid.: 116) argues, this is not how we relate to the projects we identify with and committed to. He famously asks: "[H]ow can a man, as a utilitarian agent, come to regard as one satisfaction among others, and a dispensable one, a project or attitude round which he has built his life, just because someone else's projects have so structured the causal scene that this is how the utilitarian sum comes out?"

Take *Jim and the Indians*. Jim is a harmless academic who does not want to kill: this, we can assume, is a central commitment in his life. He lives and sees this project, as it were, from the inside—from a first-person, partial point of view. However, utilitarianism is another project and maybe it is even another project of his: it is, moreover, a higher-order project that feeds on these lower-order projects since it is the satisfaction of these projects that produces the relevant valuable consequences. Now, this higher-order project requires Jim to look at his lower-order project of not killing, as it were, from the outside—from a third-person, impartial point of view.⁸ And from this viewpoint, his project loses all its peculiar, personal meaning and becomes only one project among

many other projects of many other people. It is just another project with a label on it—"permitted" or "not permitted"—and nothing more, depending on how "the utilitarian sum comes out." But this is a project that, at least in part, defines Jim as he is; it is not merely a labeled item for him.⁹ Hence utilitarianism, when it requires him to forget all about this, also alienates him from these projects and destroys the unity and shape of the particular life—*his* life—that is built around them; in short, it destroys his integrity.¹⁰

What underlies the first two problems with consequentialism is the same feature of the theory: its exclusive reliance on consequences. It is only these valuable states of affairs that matter, not how and by whom they are produced: if the right sum comes out, nothing else matters. In particular, there is no place for respecting particular persons' particular projects or welfare: people appear to be mere carriers of the good, elements in the causal chain who are needed only to produce valuable consequences. This leads the theory not to respect basic moral distinctions (how consequences are produced and who causes them) and to produce alienation and loss of integrity (given how people relate to their commitments and projects they identify with). However, as John Rawls (1971: 26–7) has famously argued, this feature also causes consequentialism to fail at the most basic metaphysical level, because it does not take seriously the separateness of persons:

The striking feature of the utilitarian view of justice is that it does not matter, except indirectly, how this sum of satisfaction is distributed among individuals any more than it matters, except indirectly, how one man distributes his satisfaction over time. The correct distribution in either case is that which yields the maximum fulfillment . . . Utilitarianism does not take seriously the distinction between persons.

The three objections paint a clear alternative picture to consequentialism. Each of us has a particular, separate life to live, a life that has a particular shape and unity, as defined by the projects and commitments we have, around which our lives are organized. Hence, it matters that we do something or someone else does or that it was our project that was thrown out in order to produce the best overall consequences.

However, while this is true, it is also clear that these objections to consequentialism are different from the demandingness objection. First, they target different things. They are not designed to question the excessive demands of consequentialism, but to object to its picture of the self (both metaphysical and motivational) and its disrespect for certain important moral distinctions. Second, the three alternative objections do not necessarily connect to excessive demands. This is clearly so with the integrity objection: any theory, whether demanding or not, violates agential integrity if it does not respect the way

agents relate to their projects and commitments. As for the other two objections, they play a part in creating excessive demands, but they certainly are not the only ingredients: maximization (of valuable consequences) is another and the particular theory of value employed (whether it concerns human or sentient welfare, for instance) is a third, and there can be others. Finally, third, one can defend the distinction between doing and allowing, argue against the separateness of persons, and find a way around the integrity objection, without doing anything about the demandingness of consequentialism.¹¹

4 Why (Only) Consequentialism?

Yet, it might be the case that the three objections are vital for understanding why the demandingness objection is typically considered as exclusively targeting consequentialism. Recall our original question: why is only consequentialism targeted in this way? It seems that other moral theories are comparably demanding. Take two popular alternatives to consequentialism. Deontology, typically associated with the name of the German philosopher Immanuel Kant, holds that there are certain things that simply cannot be done to people: these considerations, often called “rights,” function as absolute prohibitions on our actions. Another popular theory, virtue ethics, argues that the right thing to do is what the virtuous person would do, that is, the person who possesses certain character traits and dispositions (“virtues” such as honesty, courage, justice, and so on) to the maximum degree. However, as is often pointed out, Kantian prohibitions on deception can be excessively demanding in certain situations when, for instance, the only way to save lives is by lying about the whereabouts of people; emulating ideally virtuous characters (think of Mother Theresa, Gandhi, or Jesus) is arguably excessively demanding, yet, this is what virtue theory asks us to do. And these are only two examples of a non-consequentialist moral theory requiring us to give up our personal plans, projects, commitments for morality’s sake. So why is only consequentialism singled out as *objectionably* demanding? Now, one thing that could be said in response is that the three alternative objections point to those aspects of consequentialism that explain why its demandingness is objectionable: because it does not respect the separateness of persons, or because it does not respect agential integrity, or because it does not respect the distinction between doing and allowing (cf. Mulgan 2001: 15–18).

My problem with this suggestion is that it seems to go against the self-standing status of the demandingness objection. If the only problem with the demandingness of consequentialism, that is, the aspect that distinguishes its demands from the similarly excessive demands of other moral theories, is that it runs into one of the objections above, then it is these aspects—no respect for integrity, no respect for separateness, no respect for certain moral distinctions—and

not the demands *per se* that are objectionable. In other words, the reason why we reject consequentialism is not its excessive demandingness but something else that these demands only track, connected to, or derived from.

There is a clear parallel here with two influential responses to the demandingness objection. David Sobel (2007) has argued that there is a way to support the objection, but this support presupposes “prior and independent breaks with consequentialism,” that is, prior to and independent of issues of demandingness. This break concerns the distinction between the costs a moral theory requires the agent to bear and the costs a moral theory permits to befall on other people as a result of not requiring agents to prevent something happening to these people.¹² The demandingness objection, he argues, only focuses on costs a moral theory requires to bear, and totally disregards the costs a moral theory permits; this is why it says that consequentialism is objectionably demanding. However, this distinction and the resulting choice presupposes that we already know something about “the true shape of morality” before we employ the objection. That is, when we are concerned with the objection what we are concerned with is not excessive demands, but something else that our complaints only track, namely the distinction between the two kinds of costs.

Liam Murphy (2000) bases his cooperative consequentialism on an analogous diagnosis of the demandingness objection. Unlike Sobel, his focus is not on certain moral distinctions but on the fairness of moral demands. His claim is that the problem with consequentialism is not that it demands too much, but that its demands are unfair. Consequentialism is insensitive to the contributions of others, hence it requires one to contribute more if the others contribute less. In other words, it requires the agent to pick up the slack that is produced by the non-compliance of others. This indeed makes the demands of consequentialism excessive but this is not why we object to them: we object to them because they are unfair. So again, although there is an issue with the excessive demands of consequentialism, their excessiveness is not what the issue is; it at best tracks or connected to the real problem: namely, their unfairness.

It is not my aim in this chapter to evaluate these varying analyses and the responses they give rise to, but to point out that, as Murphy nicely puts it, they do not solve the demandingness objection but *dissolve* it: understood along these lines—be that integrity, separateness, fairness, or else—the demandingness objection ceases to be an objection to consequentialism on account of its demandingness. What else can we say to explain why consequentialism is objectionable solely on account of its demandingness and other, similarly demanding moral theories are not? I see three strategies. One is to refuse to answer the question. We can accept that consequentialism is not alone in being excessively demanding *and* hold that either all theories that make high demands are objectionable, or none are.¹³ The trouble with this way of reasoning is that it is widely accepted that there *is* something about consequentialism

that makes it objectionably demanding that is not present in other, similarly demanding moral theories. The best would be to respect this common intuition instead of disregarding it.¹⁴

So let us look for alternative explanations. One is offered by the notion of *supererogation*. The idea is that consequentialism, unlike non-consequentialist theories, has no place for acts that go beyond the call of duty: acts that are morally admirable, hence morally permitted, but not demanded, that is, morally required. Non-consequentialists can accept the existence of such actions since they do not require the agent to do what is best impersonally. Kantian deontology, for instance, can hold that within the bounds of certain prohibitions, actions that produce the best consequences are permitted but not required. This is why consequentialism is objectionably demanding while other theories are not: it does not leave room for these supererogatory options. The trouble with this explanation is that many question the coherence of supererogation. It appears to be paradoxical if one, reasonably, assumes that one has more moral reason to do what is better overall. Since this is the case with supererogatory acts, then the question arises: how can one be morally permitted to do something when one has *more* moral reason to do something else? There may be answers to this question, but it is also important, in the present context anyway, that the resolution of the paradox is such that it keeps the demandingness problematic in place.

Assuming the paradox can be resolved in the “proper” way, supererogation could offer us a way to explain the special demandingness of consequentialism.¹⁵ A second explanation comes from Doug Portmore (2011: 4). He focuses on what I would call consequentialism’s sensitivity to incremental changes in demands. He makes the point like this: “Utilitarianism implies that agents should sacrifice, not only their disposable income, but even their own lives and the lives of those whom they love most whenever doing so will produce the most aggregate utility, and, thus, even when the net gain would be as small as one utile” (ibid., referring to Hooker 2000: 151–2). On this explanation then, the reason why we find consequentialism objectionably demanding but other similarly demanding moral theories not, is that, unlike those other theories, consequentialism gives rise to high demands while being too sensitive to incremental changes in the overall balance of good consequences. This may well appear to be objectionably demanding, as Portmore’s quoted example illustrates.

5 Two Readings and Responses to One

The overall message of the preceding discussion is that it is not easy to find an explanation of the demandingness objection that keeps its self-standing status as well as the form in which it is normally offered: as an objection to

consequentialism and consequentialism alone. Yet, as I tried to demonstrate in the previous section, there are ways of explaining why consequentialism is the exclusive target of the objection. From this point on, therefore, I assume that the objection is in good shape. How can one respond to it? In this short section, I would like to go through the main responsive strategies (although not all of them, as I explain below). To do this, the best is to start afresh; having reached half way in the chapter, this is a good idea anyway. Take the objection again. What does it say?

We should separate two readings of the objection. They are distinguished by how one spells out the idea that excessive consequentialist demands are objectionable (Portmore 2011: 26 referring to Dorsey 2012). Consequentialism can be understood as *wrongfully* demanding if it requires agents to make sacrifices that they are not, in fact, morally required to make. Alternatively, consequentialism can be understood as *unreasonably* demanding if it requires agents to make sacrifices that they do not have decisive reason to make.¹⁶ In this and the coming two sections I focus on the first reading, which has a more influential historical pedigree.

For a (new) start then, let us pull the threads together. The demandingness objection starts from the excessive nature of consequentialist demands. It claims that these demands, therefore, are objectionable. We now know what this stands for: consequentialist demands are objectionable because they are not demands that we are in fact morally required to make. In short, consequentialism is wrongfully demanding. There is still the question why this is so, given that other moral theories make similarly excessive demands with a similarly detrimental effect on our personal, non-moral lives. Why is consequentialism the only target of the objection? This is a difficult question to answer but in the previous section I provided two—admittedly, provisional—answers: excessive consequentialist demands do not leave room for supererogatory actions, or, they are absurdly insensitive to incremental changes in the overall balance of valuable consequences. This is what makes them not only excessive, but objectionable, in the moral sense outlined here.

So this is where we stand. How can we respond to the objection understood in this way? We can put the objection somewhat more formally as making the following argument:

- (1) Consequentialism makes demand D;
- (2) Demand D is a wrongful demand;

Therefore,

- (3) Consequentialism is wrongfully demanding;
- (4) If a moral theory is wrongfully demanding, then we have reason to reject it;

Therefore,

- (5) We have reason to reject consequentialism.

This more detailed structure gives us guidance in devising the most effective response strategy. Of these, the following two responses figure most frequently in the literature.¹⁷ The *strategy of denial* rejects premise (1) either because it holds that the premise rests on false empirical facts or because it aims to restructure consequentialism in such a way that it no longer makes the demand.¹⁸ Taking an entirely different stance, the *strategy of extremism* does not deny that consequentialism makes high demands; what it denies is that these high demands are objectionable because not right: that is, it rejects premise (2).¹⁹ It does this by undermining or discrediting the intuition that the premise uses as support. Thus, it is argued that this intuition rests on lack of information, lack of clear thinking, lack of imaginative empathy or on some psychological “failure,” or that it tracks something entirely different from issues of excessive demands, or that its (typically, evolutionary) origins are such that we have no reason to take it seriously.²⁰

Much has been written about these attempts but in this chapter I will not aim to rehearse the points already made about well-known approaches. Instead, in the following two sections, I will focus on two ideas that have been left relatively unexplored by others. They can both be categorized as versions of the strategy of denial. I am not convinced about the success of either and will therefore approach them with caution: I will try to sketch how the answer to the demandingness objection would go and make some critical points along the way. First I will discuss a new version of consequentialism, called multidimensional consequentialism that was recently developed by Martin Peterson. Then I will move on to the role institutions may play in reducing the demands of consequentialism on individuals. In the last substantial section of this chapter I will turn to the second reading of the objection mentioned above in order to sketch possible responses and their potential problems.

6 Multidimensional Consequentialism and Demandingness

In his recent book, Martin Peterson (2013) puts forward a new version of consequentialism that he dubs “Multidimensional Consequentialism” (MDC).²¹ Peterson claims that his theory is in significant respects superior to other consequentialist theories: it is intuitively more appealing and it manages to avoid many of the influential objections to consequentialism, among them our particular interest, the demandingness objection.

The best place to start our investigation is Peterson’s definitions of MDC (3f.):

Let the set of C*-aspects be the set of all properties that can affect an act’s deontic status according to consequentialist theories, e.g., the wellbeing

produced by the act, or its degree of equality, and so on. The key distinction researched in this book can then be stated as follows:

One-dimensional consequentialism =_{def} the view that an act's deontic status can be characterised by a one-place function of some C*-aspect.

Multidimensional consequentialism =_{def} the view that an act's deontic status can only be characterised by a function of several C*-aspects.

Several elements of these definitions require explanation and/or further elaboration. First, in line with how I understand the theory, consequentialism on Peterson's reading is the view that "the deontic status of an act depends only on consequences" (1). Peterson dubs this principle C*. Moral aspects are those properties that directly influence the deontic status of an act, where "influence" is understood in terms of functions: "An aspect, *a*, directly influences the deontic status, *d*, of an act if and only if *d* is a function of *a*" (3).²² Putting these three—principle C*, moral aspects, influence as function—together, we get Peterson's definitions above.²³

Second, moral aspects are not the same as moral dimensions. "A dimension," Peterson explains, "can be conceived of as the conceptual space in which an aspect can be altered" (4).²⁴ This also means, as Peterson subsequently admits, that a consequentialist theory that identifies several moral aspects as affecting the deontic status of an act need not be properly speaking multidimensional because all these aspects might belong to the same dimension. However, for reasons of convenience and because his particular version will identify moral aspects that belong to different dimensions, Peterson keeps the label "multidimensional' throughout the book and I follow him on this.

Finally, third, moral aspects that determine deontic status must be irreducible. This is in fact *the* defining thesis of MDC and follows from the definition given above. Peterson labels the thesis C1. One-dimensional consequentialists must reject C1; multidimensional consequentialists must endorse it. However, Peterson goes on to claim that "in order to formulate a normatively plausible multidimensional theory, which fits well with our considered intuitions, two further non-definitional claims need to be added. Both these claims raise substantial moral issues and are logically independent of C1" (8).

The first additional claim is C2: "The binary relation 'at least as good consequences as' is not a complete ordering" (8). The idea behind this thesis is that different moral aspects are either incomparable or on a par—"that it is impossible to establish a precise exchange rate between all relevant aspects"²⁵ (9). The last defining thesis of Peterson's version of MDC is given by C3: "Moral rightness and wrongness are non-binary entities, meaning that moral rightness and wrongness vary in degrees" (9). Peterson's idea is simple: not all acts are either entirely right or wrong; some acts fall within this spectrum being in part right and in part wrong. To sum up, Peterson cashes out the deontic

status—the all-things-considered moral rightness or wrongness—of an act as a function of separate, irreducible, and incomparable (on a par) moral aspects (dimensions). He explicitly mentions three such aspects—well-being (persons), equality, and risk—but this is not intended as a comprehensive list.

There is, of course, a lot to say about the plausibility of Peterson's theory but this is beyond the scope of this chapter.²⁶ We should instead concentrate on how MDC responds to the demandingness objection. Peterson claims that MDC has the resources to defuse the objection because it can hold that donating is both right and wrong at the same time (47–8). The idea, as Peterson explains (70), is that those who donate excessively damage their own as well as their loved ones' well-being. That this is so is hard to doubt: the demandingness objection builds just on this observation (recall my introduction in section II). Since, according to Peterson, persons' well-being count separately, this influences the calculation of all-things-considered rightness by making excessive donation less right and more wrong. Why does each person's well-being count separately? Because, as I indicated above, "person" is a separate moral aspect in MDC: Unlike customary consequentialist calculations that, as we saw, do not take account of whose well-being is affected, MDC considers each person's well-being separately, as playing a separate part in determining the deontic status of an act. This helps the theory to avoid the separateness of persons objection as well as, so Peterson claims, the demandingness objection.

I find this new take on consequentialism intriguing and its response to the demandingness objection appealing. Yet, I would like to raise one critical point that offers at least something to ponder upon for an advocate of MDC. Start with the following general problem. One purpose of (deontic) all-things-considered judgments—judgments one arrives at after having taken into consideration everything that pertains to the rightness or wrongness of the given action—is the provision of *action-guidance*. However, all-things-considered judgments are not action-guiding in a satisfactory way if they do not single out at least one action as *the* thing to do—and this is exactly what MDC doesn't do. (Imagine the following conversation: "What ought I to do?"—"Well, there is nothing it would be entirely right for you to do. To some extent . . ."—"What?!") This seems to spell trouble for the theory.

Now, Peterson could reply that on his theory the thing to do is the action that is *most right* in the given circumstances.²⁷ However, if the thing to do on MDC is the act that is most right in the given situation, it is far from clear that Peterson can indeed disarm the demandingness objection. For, if the thing to do is the act with the highest deontic score ("most right"), it is well possible that, given the world as it is, consequentialism will still come out as excessively demanding. In other words, it is not enough if Peterson can show that excessive donation on MDC is not entirely right; he must also show that the ranking of alternative acts is such that excessive donation does not come out on top.²⁸ Although this is no knockdown objection to MDC, it shows that lots

of details must be filled in before we get a truly convincing response to the demandingness objection. MDC doesn't, in other words, just by construction, accomplish this.

Peterson could try to get around this problem by holding that the thing to do is not what is most right to do in the given situation but what is *sufficiently right* to do. However, besides the fact that this raises the question of where we draw the line (what is sufficiently right?), we also end up with the mirror of the debate about satisficing consequentialism (and/or sufficientarianism in theories of distributive justice).²⁹ Another possible way-out for Peterson would be to adopt agent-relative theories of value—or maybe a person-relative dimension of value. Again, however, this would leave us with the mirror version of an ongoing debate.³⁰ In general, arguments based on satisficing, agent-relative value, and so on, would be disappointing in the present context, for the hope was for MDC to escape the demandingness objection in virtue of multidimensionality and not in some other way.

7 Institutions and Demands

The core idea of this response, well known from the literature on John Rawls' theory of justice, is to direct attention to the ability of *institutions* to reduce moral demands on individuals. This is possible because a *division of labor* is justifiable: the demanding moral principles regulate institutions, whereas individuals "only" have the duty to set up and maintain these institutions. However, in order to get off the ground, this "institutional approach" has to tackle two basic challenges. First, Liam Murphy (1998) has argued that demandingness considerations will not give us what he calls *dualism*: the idea that different principles apply to institutions and to individuals. And, the thought is, we need dualism to substantiate the present response to the demandingness objection. Second, consequentialism, unlike, for instance, the Rawlsian system, appears to be a monist theory in Murphy's sense: the same principle (of beneficence) applies to individuals as to institutions. Hence the dualist idea that is taken to underlie the present response to the objection may not be justifiable in the case of consequentialism, whether or not the demandingness objection can lead us to dualism.

I believe that both objections can be answered. There is, first, the question whether we indeed need to appeal to dualism in order to respond to the demandingness objection. As Murphy's own discussion demonstrates (*ibid.*: 262–3), this need not be so: a monist theory can accommodate division of labor between institutions and individuals without making use of dualism itself. This is because it simply makes good sense, from within the monist theory, to leave the thrust of the burden of justice (Murphy's primary interest) to institutions allowing people to live their lives. Second, one can keep dualism as the

answer to the demandingness objection without going along with the stronger idea that it is the objection itself that necessitates our endorsement of dualism. Rawls and others provide good reasons in favor of dualism—I mention some of these below—that are not discussed by Murphy. Once these reasons are on the table, one can hold that we should endorse dualism for these reasons and this will still give us a response to the objection as a (perhaps unintended) side effect of the division of labor that dualism secures for us.

There is, moreover, and despite Murphy's point above, good reason to endorse dualism and not simply to rely on monism's ability to accommodate the idea of division of labor. In the (Rawlsian sense) non-ideal circumstances we live in, a monist theory poses too much risk for those who want to tackle the demandingness objection.³¹ For it is likely to be the case that in many circumstances, think of global challenges for instance, we cannot rely on institutions to do the bulk of the work for us (either because they do not exist or because they are not efficient enough). In such cases monism requires individual contribution that might well turn out to be excessively demanding. This, however, makes the second problem above even more pressing. Rawls and others following him use consequentialism as the prime example of a comprehensive, monist theory: the principle of beneficence should apply both to institutional and to individual conduct. How can we deny this? The answer is that we do not have to deny it insofar as it is properly understood. Let me explain.

The key move here is to introduce a distinction discussed at length by Samuel Scheffler (2005, 2006). There are two versions of the idea of division of labor in Rawls's work. There is, first, a division of *moral* labor that urges us to have separate moral principles for institutions and individuals on the ground that they promote different moral values. Since the relevant moral values in the case of individuals also have to do with partial concerns—such as special relationships or self-interest—this is indeed a form of division of labor that consequentialism cannot make use of; on this reading consequentialism must be a monist theory.

The *institutional* division of labor, on the other hand, relies on the idea that there are two kinds of social rules—one for the design of the basic institutional structure of society and the other for individual conduct. Principles of justice belong to the first kind for several reasons, most prominent among them is the consideration that in maintaining what Rawls calls background justice, epistemological challenges arise that cannot be faced by individuals on their own. It simply takes a lot to figure out in a complex system like a state-governed society what exactly a moral principle, even if it is simple, requires: no individual is capable of gathering the relevant data and carry out the necessary computations and reasoning. Another good reason for the institutional division of labor is the constitutive role institutions play in determining the demands of justice (Miklósi 2008 and Miklós 2013). To mention

one consideration, fundamental moral principles underdetermine moral requirements; hence, it is not possible to understand what a moral principle demands prior to the operation of institutions. One way this can happen is that moral principles, although give us a set of options to choose from, cannot make the choice themselves: they do not single out a unique set of distributive shares, or rights, or obligations. Institutions, such as the legal system, can however do just this—and until this is done, it is not determined what *exactly* is the right thing to do.

It seems to me that both considerations can also be applied to the case of consequentialism. Application of the theory clearly faces serious epistemological challenges, nor is the theory different from its main competitors concerning the indeterminacy of its requirements. If this is so, it seems we have found a way for marrying consequentialism and dualism. Moreover, if this claim is sound, it should also suffice to answer Murphy's influential objection to dualism: that it is perverse to require people to establish and maintain just (in this case: consequentialist) institutions, but not require them to personally pursue the aim of justice (when this is the most efficient way to proceed). For, there *are* good reasons to single out institutions as morally special (in fact, there are more good reasons than what I have—very briefly—presented above) that make a perfectly good case for why individuals shouldn't—because, as far as the reasons above are concerned, couldn't—pursue the aim of justice individually.³²

Having taken (very provisional) care of these initial problems, we can move on to consider the institutional response to the demandingness objection on its (substantial) merits. There are several issues that need to be discussed (including empirical questions concerning the exact demandingness of the institutional approach) but here I only focus on one that I find particularly interesting: *global justice*. Arguably, the demandingness objection is most persuasive when we appeal to existing global problems (what justice, peace, or the environment would require on the global scale). However, it might seem that the institutional approach is in trouble here since the relevant institutions, but not the demands are missing; hence, dualism cannot be appealed to in response to the objection in this case.³³ One reply to this objection is to endorse what is often called the *relationalist* position in the literature on global justice: that claims of justice are grounded in certain institutional relations among people, such as, to mention another influential Rawlsian thought, the mutually beneficial cooperative relations people often maintain. Hence the response: since these relations do not exist globally, there are also no global moral demands. However, I am not personally inclined to endorse this way of thinking about global justice; besides, and this is more important in the present context, consequentialism is the prime example of a *non-relationalist* theory, that is, one that does not ground claims of justice in institutional relations among people.³⁴ Consequentialists seem to be committed to the thought that we have moral duties in virtue of our

common humanity, for example, the fact, that we all can feel pain and pleasure, or be well-off or badly off in some other way.

With the relationist approach out of the way, we need to find the institutions that can be used as *instruments* to carry out (and in part constitute) what consequentialism requires on the global level. Without this we cannot make our dualist approach work in practice. Can we find the relevant institutions? There are two ways to proceed. One is to point to already existing institutions on the global level; this is what relationalist advocates of global justice do and we can borrow from them at this point.³⁵ Here one can cite such examples as the World Trade Organization (WTO), the International Monetary Fund (IMF), or the World Bank but there is a lot of empirical research done in this field that we cannot do justice here.³⁶ The point is that there *are* already several institutions that can be used for the purposes of fulfilling consequentialist requirements. Naturally, a lot more can be done to improve these institutions and it is a largely empirical matter how this will look like and what it will require (and how demanding this will be).

Another way to go about responding is to make a radical break with what we can consider to be the *status quo*: why not build a global state instead of relying on already existing but rather constrained institutions? This is what Torbjörn Tännsjö (2008) suggests that we should do. He argues that the three major global problems—lack of world peace, environmental problems, and problems of justice—can only be tackled by a world state. Moreover, he adds, we have a unique window of opportunity to build such a state: the fact that we have only one superpower in existence, namely, the United States. Tännsjö then goes on to master empirical as well as theoretical support for these claims, arguments that I cannot do justice here. However, the message is clear. Global moral demands are real and cannot be evaded. Moreover, we either already have the means for tackling them, or we can develop these means—if needed, in the form of a world state.

To sum up, the institutional approach to the demandingness objection is a promising but certainly insufficiently worked out way to respond to the challenge. The problems are both theoretical and empirical in nature, as I have attempted to demonstrate above; yet, I believe it is worth the effort to work out this approach in detail to see where it takes us and what we can achieve with it.

8 Consequentialist Reasons and Demandingness

We are nearing the end of a long journey. There is one response left to discuss but to do this, I need to return to the second reading of the objection in section V and say a bit more about how it unfolds. Recall, on the second reading the objection claims that consequentialism is *unreasonably* demanding: it

requires us to do things that we do not have decisive reason to do. We can formalize the argument in the following way:

- (1) Consequentialism makes demand D;
- (2) Demand D is unreasonable;

Therefore,

- (3) Consequentialism is unreasonably demanding;
- (4) If a moral theory is unreasonably demanding, then we have reason to reject it;

Therefore,

- (5) We have reason to reject consequentialism.

Unlike the analogous reasoning of the first reading, this argument requires further elaboration. Premise (2) is again supported by an intuition that we supposedly share. It is that consequentialist reasons are not the only reasons around and at least some of the alternative reasons are stronger than consequentialist reasons. Portmore (2011: 32) identifies two such classes of reasons:

- (1) reasons that have nothing to do with promoting the good, such as the reason one has to refrain from violating someone's autonomy even when doing so is a means to promoting the good, and (2) reasons that stem from the special relations that we bear to ourselves and our loved ones, such as the reason one has to promote the good by saving one's own loved one as opposed to by helping some stranger save her loved one.

Some of these reasons are moral (those rooted in our respect for others' autonomy or those grounded in special obligations), others are non-moral (those grounded in our self-interest), but their common feature is that consequentialism, as understood here, cannot accommodate them: they have nothing to do with the maximization of impersonal goodness.

The other critical point of the argument is premise (4). Unlike the analogous premise of the first reading, the truth of this premise is not obvious. It is based on the thesis that what morality requires us to do must also be rationally authoritative: it must be backed by decisive reasons. *Moral rationalism*, as Portmore (ibid.: 28) calls the thesis, brings together the moral and the rational and has many supporters. They understand moral rationalism as a constraint on moral theories: if a moral theory turns out to make demands on us that we are not rationally required to fulfill, it is not a defensible moral theory. Consequentialism is a case at hand due to the existence of (sometimes) stronger non-consequentialist reasons.³⁷

Notice, moreover, that this account of why we should reject consequentialism could give us a third explanation why the theory is the sole target of the demandingness objection: unlike other moral theories, consequentialism does not respect the existence of non-consequentialist reasons. This is up for discussion, though (do other moral theories indeed respect *all* these reasons?); but even if this claim turns out to be not the case, the present story can well supplement the two explanations mentioned in section IV. Portmore (*ibid.*: 4) is clear about this. Consequentialism's sensitivity to incremental changes makes it particularly liable to encounter situations in which non-consequentialist reasons come out as winners: it is hard to accept that a tiny little improvement in the overall goodness of consequences would be enough to rationally justify acting against (some perhaps quite powerful) non-consequentialist reasons.

Let us return to the argument. Counterattacks can be launched at three points.³⁸ Premise (1) can again be rejected. This is typically argued by showing that consequentialism can accommodate the kinds of reasons mentioned by Portmore. These attempts have received sufficient attention in the literature and I will not discuss them here.³⁹ Next, one can reject premise (2). This can again be done in extremist fashion as in the case of the first reading of the objection. Alternatively, one can carry out empirical research to see if the premise is indeed intuitively supported: if it is indeed part of commonsense morality that these reasons exist and are (sometimes) stronger than consequentialist reasons. However, although I am supportive of such investigations, I certainly would not consider their results decisive.⁴⁰

A third line of response rejects premise (4) by giving up moral rationalism. On the face of it, the attempt is doomed to failure because it marginalizes morality. This is how Hurley (2009: 60) puts the problem:

If we accept that morality, properly understood, provides merely one among other sets of standards, and that this set of standards lacks the distinctive relationship that has been claimed for it to our reasons for acting, then morality is shifted toward the margins of meaningful inquiry into what rational agents such as ourselves have reasons to do. This would be a pyrrhic victory for the consequentialist, vindicating his account of moral standards only by marginalizing the role of such standards in practical reason and deliberation.

I have two problems with Hurley's marginalization charge. One, it is not obviously counterintuitive to hold that consequentialist reasons do not always outweigh our non-consequentialist reasons. Portmore's point about the existence and strength of these alternative reasons suggests this much. So it is far from clear to me why, *contra* Hurley, consequentialists shouldn't embrace the

denial of moral rationalism: *consequentialist* morality is often a marginal affair from a rational viewpoint, why should we deny this? Second, consequentialism, as usually conceived, is a theory of moral standards and not a theory of moral reasons: it is a thesis on what is the right or wrong thing to do, not a claim about what we have reason to do. Hence, we have to augment consequentialism with a suitable theory of practical reasons and then see if this is in line with moral rationalism. However, this is clearly an open-ended project that shouldn't be prematurely given up just because one insists on the truth of moral rationalism.⁴¹

So far we have operated with a picture of morality and rationality that takes them to be separate realms. One has a theory of moral standards and one has a theory of reasons and then combines them in some way to see if we get to moral rationalism. This way of proceeding seems to equate moral rationalism with what is often called the *overridingness thesis*: the claim that moral reasons and requirements override non-moral reasons and requirements (Scheffler 1992: 52; Stroud 1998: 171). That the two doctrines amount to the same thing is suggested by the following line of reasoning. Let's first suppose that reasons determine what we are morally required to do. Reasons are not added to or derived from an already existing moral picture, but they are the primary determinants of moral requirements. Let's further suppose that in determining what we are morally required to do, *moral* reasons play a decisive role: they are the only reasons that count in determining our moral requirements. Finally, let's suppose moral rationalism is true, and hence that what morality requires us to do must also be rationally authoritative. The end result is the overridingness thesis: moral reasons and requirements will invariably outweigh all *non*-moral reasons and requirements.

But, as Portmore (2011: 38–40) points out, it is a mistake simply to equate moral rationalism with the overridingness thesis. In fact, disambiguating these theses is crucial because it shows us that there is another way for consequentialists to avoid the demandingness objection. Portmore embraces this opportunity. He argues that, contrary to the second claim above, moral *as well as* non-moral reasons determine what we are morally required to do. Why is this crucial? Because it makes the following strategy possible (ibid.: 41–2):

So although some may be compelled to accept moral rationalism because they think that moral requirements generate overriding reasons to abide by them, others, like myself, may be driven to accept moral rationalism because they think that morality is limited in what it can require of us—that morality can require us to do only that which we have decisive reason to do, all things considered. The thought would be that although moral requirements do not generate overriding reasons to abide by them, moral rationalism is, nevertheless, true, for non-moral reasons

serve to constrain what morality can require of us in that they sometimes successfully counter our moral reasons, preventing them from generating moral requirements.

Schematically, then, Portmore argues like this. One, reasons determine moral requirements; two, these reasons are moral and non-moral, consequentialist as well as non-consequentialist; three, the morality so produced is nonetheless a consequentialist morality. This reasoning indeed preserves moral rationalism since, as Portmore above explains, nothing will be allowed to come out as a moral requirement that doesn't have the support of reasons.

Of course, the reasoning is eminently questionable.⁴² I already mentioned that step one offers only one picture of morality and rationality and I added that Portmore argues for step two. It is an even less straightforward matter, and takes Portmore the rest of his book to argue for, that the resulting morality will be consequentialist.⁴³ Finally, and perhaps most importantly, Portmore's consequentialism is not the consequentialism of this chapter. He takes consequentialism of that kind to be defeated by the demandingness objection. In other words, what Portmore does is to restructure consequentialism and thereby deny premise (1) of the demandingness objection. Hence, even if his argument succeeds, there will still be the question whether his consequentialism is indeed one we want to accept. Needless to say, just as with the other two responses, there is no space here to take up any of these matters. I believe Portmore's theory is promising and deserves detailed discussion, but this cannot be conducted in this chapter.⁴⁴

9 Summary and Concluding Remarks

In this chapter I set out to do three things. (1) I wanted to clarify what the demandingness objection is about. (2) I wanted to explain why the objection targets only consequentialism. (3) I wanted to present responses that are (relatively) underexplored in the literature. I hope to have accomplished the first two tasks and at least partially carried out the third. I do not claim to have not left questions open; in fact, at least in the case of potential responses, my intention *was* to raise questions. To answer those questions, however, is a task left for another occasion.

Notes

- 1 Those who want to read about consequentialism in more detail can turn to Sinnott-Armstrong (2014). Mulgan (2007) and Bykvist (2010) also provide very good introductory discussions and of course most introductory ethics texts will have a detailed discussion of the theory.

- 2 Henceforth: “consequentialism,” unless qualifier needed. In the rest of this chapter I will use the terms “consequentialism” and “utilitarianism” synonymously: the objections to utilitarianism that I will consider are also objections to consequentialism and vice versa.
- 3 The first comes from Mulgan (2007), page 95, the second from Bykvist (2010), page 98.
- 4 See Cullity (2004) and (2009) for introducing and making substantial use of these iterated demands.
- 5 Precision at this point would require clarification of the notion of “demand,” but I cannot do this in the present chapter. I wrote more about this in Tanyi (2012). One point deserves short notice, though. Some philosophers argue that in addition to costs one should also consider the factor Scheffler (1992: 98) calls *confinement*. In his formulation a moral theory is confining to the extent the constraints it involves narrow the range of morally acceptable courses of action open to the agent. It is, however, questionable whether confinement indeed constitutes an independent factor. Murphy (2000: 29–30), for instance, argues that a large part of confinement can be explained as losses that the agent suffers in her well-being when obeying with moral dictates, and this is just the traditional understanding of a moral demand.
- 6 Unfortunately, it is easy to cite statistics for this claim. Any report by the World Health Organization (WHO), the World Bank, the United Nations Children’s Fund (UNICEF), the United Nations Development Programme (UNDP), and so on paints the same dire picture, certainly of the global situation, but also, in most cases, of domestic circumstances. See Miklós (2013: 2–3) for more data and references.
- 7 See also Williams (1973b) and Stocker (1976) for a very similar objection. It is clear that the integrity objection is closely allied to Williams’ thinking about internal reasons in Williams (1981a). See Hurley (2009: Chapter 4) who works out this connection in detail.
- 8 What happens if the agent has *no* lower-order projects? In this case, no alienation and integrity charge follows. However, this is certainly an extreme case, the truest form of what Wolf (1982) calls a moral saint, and comes with other problems, as Wolf demonstrates. Moreover, it is also clear that not everyone can be (even if, *contra* Wolf, should be) a moral saint. Being a higher-order project, utilitarianism needs lower-order projects, that is, it needs people who have those projects and are thus not saints themselves.
- 9 Hence Williams’ (1981b) charge that utilitarianism requires the agent to have “one thought too many” when doing the morally right thing. Jim’s motivating thought in helping the Indians wouldn’t merely be, as it should be, that they are in trouble and need help, but also that this action is permitted (or required) by morality. See also Smith’s (1995) related fetishism charge for further discussion.
- 10 There are some loose ends here, though. In particular, it is a question how important the utilitarian project must be as compared to the other, lower- and higher-order projects of the agent. Hurley (2009: Chapter 4) argues that it must be the agent’s most important, ultimate project that subsumes all other lower-order projects. Now, the integrity objection is about how utilitarianism forces the agent to look at his or her central commitments in life and as such, it seems to hold even if the utilitarian higher-order project is not a project of ultimate importance. However, if it is not such an ultimate project, it means that on certain occasions at least, the agent can experience his or her lower-order projects in the proper (i.e. non-alienating) way and this certainly tones down (if not eliminates) the force of Williams’ objection.

- 11 For the relevant consequentialist responses on doing versus allowing, see Howard-Snyder (2011); on the separateness of persons, see Parfit (1984), Brink (1997); on issues of integrity, see Railton (1984) and Scheffler (1992).
- 12 To give an example, in *Jim and the Indians*, there is the cost that Jim is required to bear if he decides to shoot the one chosen Indian: namely, that he becomes a killer. At the same time, there is also the cost that will befall upon all the Indians who will be shot by Pedro if Jim decides not to shoot the one chosen Indian and thereby prevent Pedro from killing every Indian.
- 13 The first option in the disjunctive statement keeps the demandingness objection, but extends it to other theories. However, the second option is in effect a response to the objection, often called the “companions in guilt” strategy. See Ashford (2003) for employing this line of reasoning on the ground that Scanlon’s contractualism is also excessively demanding. See Mulgan (2007) and Bykvist (2010) for a general discussion of this strategy.
- 14 Unless it can be proven that no such intuition really exists or for some reason it is not to be relied upon. Such result could follow if, as Parfit (2011) claims, all the main moral theories (when properly understood) converge. But of course this is a big “if” and many do not seem to agree with Parfit’s conclusion in his monumental work.
- 15 See Bykvist (2010: 105) and Portmore (2011: 134) for two attempts to resolve the paradox but not in the “proper” way, that is, without putting an end to the demandingness problem.
- 16 I understand “decisive reason” and “most reason” to mean the same. There are some complications here, which I am disregarding in this chapter. In particular, the two notions come apart if reasons can behave “unusually,” for example, if they can silence or bracket each other. For more on this kind of behavior of reasons, see Tanyi (2013).
- 17 The labels I use for these response strategies come from Mulgan (2001).
- 18 For a thorough discussion and criticism of the first, *empirical strategy*, see Mulgan (2001) and Bykvist (2010). The second approach, the *restructuring strategy* has given rise to such positions as sub-maximizing consequentialism (Slote 1984; for criticism, see Pettit 1984 and Bradley 2006), two-level consequentialism (Hare 1981; and in general the distinction between decision procedure and criterion of rightness in Railton 1984; for criticism, see McNaughton 1988), rule-consequentialism (most recently Hooker 2000; for criticism, see Mulgan 2001), limited consequentialism (Scheffler 1992, 1994; for criticism, see Kagan 1984), cooperative consequentialism (Murphy 2000; for criticism, see Mulgan 2001), and combined consequentialism (Mulgan 2001).
- 19 This is anyway what they *should* claim. Often, however, as Portmore (2011: 26) points out, the extremist claim seems to be that consequentialism is indeed too demanding in this sense, but this is not objectionable. In other words, the strategy would then accept premise (2), but deny premise (4). But this is certainly no way to proceed since consequentialists simply cannot grant to their opponents the truth of the claim that their theory makes demands that are not morally right. This would obviously be self-defeating for the theory.
- 20 See Kagan (1989); Singer (1972); Tännsjö (2002); Unger (1996); Sobel (2007); for critical discussion of this approach, see Cullity (1994) and Mulgan (2001, 2007). The intuition extremists focus on can be just those particular intuitions that certain counterexamples are meant to evoke against their theory (such as, e.g., in *Your Money and Charity*); or they can focus on moral intuitions in general. The latter way of arguing gave rise to a separate debate about the use of intuitions in moral theory. See Singer (2005) versus Sandberg and Juth (2011).
- 21 In this section, unless otherwise stated, all page references in brackets will be to this book.

- 22 That is, as Peterson elsewhere explains, “something counts as an aspect if and only the deontic status of an act varies if we hold constant everything but the putative aspect in question” (15).
- 23 The clause “can be characterized” in the definition of one-dimensional consequentialism is important because it makes the set of one-dimensional theories less restricted. For all it requires is that we find a moral aspect that makes it possible to characterize an act’s deontic status as a function of that one aspect; this does not rule out that another characterization exists that employs several moral aspects.
- 24 He brings geometry as an analogous case: “The area of the circle depends on only one aspect (its radius) whereas the area of the triangle depends on two aspects (its base and height). All three aspects are elements of the same dimension (length). This is not always the case, however, as can be seen by considering an analogy with physics: mass and time are different aspects, but they are also elements of different dimensions” (4).
- 25 Peterson defines incomparability as the claim that “for some consequences, no pairwise evaluative comparisons can be made.” As for the other notion, his definition is that “two elements are on a par if and only if they are comparable, although it is false that one is at least as good as the other” (9).
- 26 I do this in a forthcoming *Philosophical Studies* paper coauthored with Vuko Andric. We focus on the relation between C1 and C3 and the argument for C3.
- 27 Or least wrong: Since we are dealing with all-things-considered rightness and wrongness, the act that is most right is also the least wrong (the two degrees must add up to 1).
- 28 Another problem is that it is not clear that what Peterson is talking about in discussing the case is indeed degrees of rightness/wrongness and not what he calls moral *strength* (see his Chapter 2.4 on this distinction). Roughly, *how much* we donate appears to be about how much moral value we produce or fail to produce and this is strength, not degree (that is given by the important moral value of helping/saving lives, which is the same in all instances), according to Peterson (cf. 2013: 93, 117).
- 29 Satisficing consequentialism is the view that the right thing to do is what is good enough, for example, produces enough utility. See Slote (1984) for the original formulation and Bradley (2006) for criticism. Sufficientarianism is a view about how to distribute goods fairly and holds that the right distribution is the one in which no individual is allowed to fall below a certain threshold. See Crisp (2003) for a recent statement and Temkin (2003) for criticism.
- 30 Agent-relative values are values that make essential reference to the agent who has the value: the value of special relationships (family, love), for instance, is agent-relative since it is essential to mention whose relations we are talking about in order to understand the goodness of these relationships. On the agent-relative versus agent-neutral distinction in more detail, see Ridge (2011). Consequentialism traditionally operates with agent-neutral values only: it does not matter, for example, whose pleasure or pain we are talking about—the badness of these states can be understood without reference to the agent who has them. However, some consequentialists recently attempted to broaden consequentialism to accommodate agent-relative values as well. On an insightful critique of what is often called evaluator-relative consequentialism, see Schroeder (2007). A person-relative dimension of value, I take it, could be Peterson’s versions of this idea: it would incorporate agent-relativity into the dimension of person (and thus well-being).
- 31 For a good discussion of the ideal versus non-ideal theory distinction, see Simmons (2010).
- 32 Although there are other objections to dualism in the literature (I have in mind Cohen 1992, 1997, 2000 and Nagel 1991), these are discussed and responded to by

- others (such as Scheffler 2005, 2006). The chapter by Andres Moles in this volume also has a good discussion of several of the issues these authors raise.
- 33 Again, a good discussion of the upcoming issues can be found in the chapter by Andres Moles in this volume.
- 34 Some, like Nagel (2005), seem to hold that a non-relational theory must be monist, but I fail to see the connection. The relational/non-relational distinction concerns the *grounds* of justice (with consequences for its *scope*), whereas the monism/dualism distinction is about the *site* of justice. Although both invoke institutions, they do so in an entirely different role.
- 35 For an early representative see Pogge (1994); for a more recent contributions, see Moellendorf (2011).
- 36 See, for example, Nussbaum (2007) for a long list of the relevant institutions and schemes.
- 37 Hurley (2009: Chapter 2) presents this problem as a “troubling normative triad” for consequentialism. The elements of the triad are the consequentialist moral standard (CMS), the existence of non-impersonal reasons (NIR), and the rational authority of moral standards (RAMS). The point is the same as in my presentation: the elements of the triad cannot be fitted together into one coherent, defensible whole.
- 38 There are also those responses that reject the argument by questioning what it assumes. Scalar-consequentialists claim that consequentialism makes no demands on us, although it does give us reasons to act. See Norcross (2006); for a response, see McElwee (2011). What we might call normative relativism argues that every reason is relative to a point of view, hence there is common platform on which they can compete with each other. See Kagan (1989) and Copp (1997); for a criticism, see McLeod (2001).
- 39 Evaluator-relative consequentialism, mentioned in footnote 29, is one attempt, Cummiskey (1996)’s Kantian consequentialism is another. For a critical discussion of the latter see Hurley (2009: Chapter 6.2). But we should note that these theories only deal with a subset of all the “missing” reasons and hence cannot be taken as providing a complete response, even if they do not fail otherwise.
- 40 With Martin Bruder, I have been carrying out research along these lines with promising results that show that the majority of people regard consequentialist reasons as overriding other considerations. See, for example, Bruder and Tanyi (2014).
- 41 Sobel (2007: 14–15) appears to be in agreement with this. Portmore (2011) disagrees but, as we shall see, he doesn’t share this picture of morality and rationality. He also provides an intricate argument for moral rationalism that he defends in detail. I say more about his argument for moral rationalism in Tanyi (2012), but no doubt, more critical discussion would be needed.
- 42 Hurley (1999: 122), for instance, endorses Portmore’s general picture about reasons and morality—he calls it the moral authority of rational standards (MARS)—but thinks that it leads nowhere near to consequentialism.
- 43 In particular, Portmore (2011: Chapter 3) argues for a teleological conception of reasons that he claims to be embedded in consequentialism.
- 44 For some initial attempts, see Hurley (2014), Archer (2014), and Gert (2014).

References

- Archer, A. (2014). “Moral Rationalism without Overridingness.” *Ratio* 27: 100–114.
- Ashford, E. (2003). “The Demandingness of Scanlon’s Contractualism.” *Ethics* 113: 273–304.
- Bradley, B. (2006). “Against Satisficing Consequentialism.” *Utilitas* 18: 97–108.

- Brink, D. O. (1997). "Kantian Rationalism: Inescapability, Authority, and Supremacy." In G. Cullity and B. Gaut (eds), *Ethics and Practical Reason*, pp. 255–91. New York: Oxford University Press.
- Bruder, M. and Tanyi, A. (2014). "Overdemanding Consequentialism? An Experimental Approach." *Utilitas* 26 (3): 250–75.
- Bykvist, K. (2010). *Understanding Utilitarianism*. London: Continuum.
- Carter, A. (2009). "Is Utilitarian Morality Necessarily too Demanding?" In T. Chappell (ed.), *The Problem of Moral Demandingness: New Philosophical Essays*, pp. 163–185. London: Palgrave MacMillan.
- Cohen, G. A. (1992). "Incentives, Inequality, and Community." In G. B. Peterson (ed.), *The Tanner Lectures on Human Values*, Vol. 13. Salt Lake City: University of Utah Press.
- Cohen, G. A. (1997). "Where the Action Is: On the Site of Distributive Justice." *Philosophy and Public Affairs* 26 (1): 3–30.
- Cohen, G. A. (2000). *If You're an Egalitarian, How Come You're So Rich?*. Oxford: Oxford University Press.
- Copp, D. (1997). "The Ring of Gyges: Overridingness and the Unity of Reason." *Social Philosophy and Policy* 14: 86–106.
- Crisp, R. (2003). "Equality, Priority, and Compassion." *Ethics* 113: 745–63.
- Cullity, G. (1994). "International Aid and the Scope of Kindness." *Ethics* 105: 99–127.
- Cullity, G. (2004). *The Moral Demands of Affluence*. Oxford: Clarendon Press.
- Cullity, G. (2009). "Demandingness and Arguments from Presupposition." In T. Chappell (ed.), *The Problem of Moral Demandingness: New Philosophical Essays*, pp. 8–35. London: Palgrave MacMillan.
- Cummiskey, D. (1996). *Kantian Consequentialism*. Oxford: Oxford University Press.
- Dorsey, D. (2012). "Weak Anti-Rationalism and the Demands of Morality." *Nous* 46 (1): 1–23.
- Gert, J. (2014). "Perform a Justified Option," *Utilitas* 26 (2): 206–17.
- Hare, R. M. (1981). *Moral Thinking*. Oxford: Clarendon Press.
- Hooker, B. (2000). *Ideal Code, Real World: A Rule Consequentialist Theory of Morality*. Oxford: Clarendon Press.
- Hooker, B. (2009). "The Demandingness Objection." In T. Chappell (ed.), *The Problem of Moral Demandingness: New Philosophical Essays*, pp. 148–63. London: Palgrave MacMillan..
- Howard-Snyder, F. (2011). "Doing vs. Allowing Harm." *The Stanford Encyclopedia of Philosophy* (Winter 2011 Edition), Edward N. Zalta (ed.), <<http://plato.stanford.edu/archives/win2011/entries/doing-allowing/>>.
- Hurley, P. (2009). *Beyond Consequentialism*. New York: Oxford University Press.
- Hurley, P. (2014). "Comments on Douglas Portmore's *Commonsense Consequentialism*." *Philosophy and Phenomenological Research* 88 (1): 225–32.
- Kagan, S. (1984). "Does Consequentialism Demand Too Much?" *Philosophy and Public Affairs* 13 (3): 239–54.
- Kagan, S. (1989). *The Limits of Morality*. Oxford: Clarendon Press.
- Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong*. London: Penguin.
- McElwee, B. (2011). "Impartial Reasons, Moral Demands." *Ethical Theory and Moral Practice* 14: 457–66.

- McLeod, O. (2001). "Just Plain 'Ought.'" *Journal of Ethics* 5: 269–91.
- McNaughton, D. (1988). *Moral Vision*. London: Wiley-Blackwell.
- Miklós, A. (2013). *Institutions in Global Distributive Justice*. Edinburgh: Edinburgh University Press.
- Miklósi, Z. (2008). "Compliance with Just Institutions." *Social Theory and Practice* 34 (2): 183–207.
- Moellendorf, D. (2011). "Cosmopolitanism and Compatriot Duties." *The Monist* 94 (4): 535–54.
- Mulgan, T. (2001). *The Demands of Consequentialism*. Oxford: Clarendon Press.
- Mulgan, T. (2007). *Understanding Utilitarianism*. Stocksfield: Acumen.
- Murphy, L. D. (1998). "Institutions and the Demands of Justice." *Philosophy and Public Affairs* 27 (4): 251–91.
- Murphy, L. D. (2000). *Moral Demands in Nonideal Theory*. Oxford: Oxford University Press.
- Nagel, T. (1991). *Equality and Partiality*. New York: Oxford University Press.
- Nagel, T. (2005). "The Problem of Global Justice." *Philosophy and Public Affairs* 33 (2): 113–47.
- Norcross, A. (2006). "Reasons without Demands: Rethinking Rightness." In James Dreier (ed.), *Contemporary Debates in Moral Theory*, pp. 38–53. Oxford: Blackwell.
- Nussbaum, M. (2007). *Frontiers of Justice*. Cambridge, MA: Harvard University Press.
- Parfit, D. (1984). *Reasons and Persons*. Oxford: Oxford University Press.
- Parfit, D. (1984). *On What Matters*. Oxford: Oxford University Press.
- Peterson, M. (2013). *The Dimensions of Consequentialism*. Cambridge: Cambridge University Press.
- Pettit, P. (1984). "Satisficing Consequentialism." *Proceedings of the Aristotelian Society*, Supplementary Volume 58: 165–76.
- Pogge, T. (1994). "An Egalitarian Law of Peoples." *Philosophy and Public Affairs* 23 (3): 195–224.
- Portmore, D. W. (2011). *Commonsense Consequentialism: Wherein Morality Meets Rationality*. Oxford: Oxford University Press.
- Railton, P. (1984). "Alienation, Consequentialism, and the Demands of Morality." *Philosophy and Public Affairs* 13: 134–71.
- Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.
- Ridge, M. (2011). "Reasons for Action: Agent-Neutral vs. Agent-Relative." *The Stanford Encyclopedia of Philosophy* (Winter 2011 Edition), Edward N. Zalta (ed.), <<http://plato.stanford.edu/archives/win2011/entries/reasons-agent/>>.
- Sandberg, J., and Juth, N. (2011). "Ethics and Intuitions: A Reply to Singer." *Journal of Ethics*, 15 (3): 209–26.
- Scheffler, S. (2005). "The Division of Moral Labour: Egalitarian Liberalism as Moral Pluralism." *Proceedings of the Aristotelian Society*, Supplementary Volume 79: 229–53.
- Scheffler, S. (2006). "Is the Basic Structure Basic?" In C. Sypnowich (ed.), *The Egalitarian Conscience: Essays in Honour of G.A. Cohen*, pp. 102–29. Oxford: Oxford University Press.
- Scheffler, S. (1992). *Human Morality*. New York: Oxford University Press.

- Scheffler, S. (1994). *The Rejection of Consequentialism*. Oxford: Clarendon Press, Revised edition.
- Schroeder, M. (2007). "Teleology, Agent-Relative Value, and 'Good'*." *Ethics* 117: 265–95.
- Sidgwick, H. (1907). *The Methods of Ethics*. Cambridge: Cambridge University Press, 7th edition.
- Simmons, A. J. (2010). "Ideal and Nonideal Theory." *Philosophy and Public Affairs* 38 (1): 5–36.
- Singer, P. (1972). "Famine, Affluence, and Morality." *Philosophy and Public Affairs* 1 (3): 229–43.
- Singer, P. (2005). "Ethics and Intuitions." *Journal of Ethics* 9: 331–52.
- Sinnott-Armstrong, W. (2014). "Consequentialism." *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), Edward N. Zalta (ed.), <<http://plato.stanford.edu/archives/spr2014/entries/consequentialism/>>.
- Slote, M. (1984). "Satisficing Consequentialism." *Proceedings of the Aristotelian Society* 58: 139–63.
- Smith, M. (1995). *The Moral Problem*. New York: Oxford University Press.
- Sobel, D. (2007). "The Impotence of the Demandingness Objection." *Philosophers' Imprint* 7: 1–17.
- Stocker, M. (1976). "The Schizophrenia of Modern Ethical Theories." *Journal of Philosophy* 73: 453–66.
- Stroud, S. (1998). "Overridingness and Moral Theory." *Pacific Philosophical Quarterly* 79: 170–89.
- Tännsjö, T. (2002). *Understanding Ethics: An Introduction to Moral Theory*. Edinburgh: Edinburgh University Press.
- Tännsjö, T. (2008). *Global Democracy: The Case for a World Government*. Edinburgh: Edinburgh University Press.
- Tanyi, A. (2012). "The Case for Authority." In S. Schleidgen (ed.), *Should We Always Act Morally? Essays on Overridingness*, pp. 159–89. Marburg: Tectum.
- Tanyi, A. (2013). "Silencing Desires?" *Philosophia* 41 (3): 887–903.
- Temkin, L. (2003). "Egalitarianism Defended." *Ethics* 113: 764–82.
- Unger, P. (1996). *Living High and Letting Die*. New York: Oxford University Press.
- Williams, B. (1973a). "A Critique of Utilitarianism." In J. J. Smart and B. Williams (eds), *Utilitarianism: For and Against*, pp. 77–151. Cambridge: Cambridge University Press.
- Williams, B. (1973b). "Morality and the Emotions." In B. Williams (ed.), *Problems of the Self*, pp. 207–30. Cambridge: Cambridge University Press.
- Williams, B. (1981a). "Persons, Character and Morality." In B. Williams (ed.), *Moral Luck*, pp. 1–19. Cambridge: Cambridge University Press.
- Williams, B. (1981b). "Internal and External Reasons." In B. Williams (ed.), *Moral Luck*, pp. 101–14. Cambridge: Cambridge University Press.
- Wolf, S. (1982). "Moral Saints." *Journal of Philosophy* 79: 419–39.

30 Political Obligation, and the Site and Scope of Justice

Andres Moles

It is impossible to overestimate the impact that the publication of John Rawls' *A Theory of Justice* has had in political philosophy in the last 40 years. Since its publication, questions of justice have dominated analytical political philosophy. Rawls' work is impressive in that it combines the values of liberty and equality in a unified conception of justice. Before Rawls, it was commonly held that these values were incompatible. Liberty could only be secured by restricting equality, and the price of equality was liberty. By contrast, Rawls argues that the most reasonable conception of justice is liberal egalitarian and consists of two principles:

- (a) Each person has an equal claim to a fully adequate scheme of basic rights and liberties, which scheme is compatible with the same scheme for all; and in this scheme the equal political liberties, and only those liberties, are to be guaranteed their fair value.
- (b) Social and economic inequalities are to satisfy two conditions: first, they are to be attached to positions and offices open to all under conditions of fair equality of opportunity; and second, they are to be to the greatest benefit of the least advantaged members of society (Rawls 1996, 5–6).

Both principles are controversial in many ways, and there is by now, a vast bibliography that attempts to refute, clarify, or defend this conception of justice and its underlying assumptions. Let me briefly mention two highly influential debates. First, in order to identify who are the least advantaged members of society, we need an answer to the question about the “currency” of justice. Moreover, given liberal egalitarians care about equality it is crucial to know in what respects people should be equal. It is far from obvious whether we should equalize access to “primary goods” (Rawls 1999a), or opportunity for welfare (Arneson 1989), or people's ability to engage in some worthwhile

activities (Sen 2000), or their resources (Dworkin 2002), or a combination of them (Cohen 2011). Second, some authors worry that Rawls' account of justice is insufficiently sensitive to disadvantages for which persons can be held responsible. It seems intuitively unfair that some people are worse-off than others through no fault of their own. At the same time, we might think that it is not objectionable that some are disadvantaged provided that they are responsible for it. For instance, it is unfair that someone is at a disadvantage because of his/her poor social background or lack of productive endowments. But it might not be unfair that someone is worse-off than others when he/she made a gamble and was unlucky. There is a vivid debate about the merits and problems of this "luck egalitarian" view (Cohen 2011 and Dworkin 2002 defend a form of luck egalitarianism, Hurley 2003, and Anderson 1999 present each sophisticated criticisms).

A different reaction to Rawls' book focused not on the content and justification of his principles of justice, but on the idea that justice has priority over questions about the good life. Communitarians argued that Rawls pays insufficient attention to the cultural background and the shared values upon which just institutions function. More recently, liberal perfectionists have defended the view that the state can legitimately promote values other than justice (Mulhall and Swift 2006 offer a superb reconstruction of this debate). Rawls addresses these worries in *Political Liberalism* (1996). According to him, it is to be expected that, under conditions of freedom, people will reasonably disagree widely about their doctrines of the good life. For instance, he thinks that there will be reasonable disagreement about religious questions, about the value of life, nature of value, and other philosophical and metaphysical issues. He calls this phenomenon the fact of "reasonable pluralism," and asks how it is possible that people who disagree deeply about the good can nevertheless be expected to endorse liberal democratic regimes. His reply is that "our exercise of political power is fully proper when it is exercised in accordance with a constitution, the essentials of which all citizens as free and equal may be reasonably expected to endorse in the light of principles and ideals acceptable to their common human reason" (Rawls 1996, 137). Moreover, he argues that all reasonable doctrines of the good endorse some fundamental *political* values: those of society as fair system of social cooperation between free and equal citizens. Endorsement of these values is what makes doctrines reasonable (Quong 2011). But there is no other common ground that everyone accepts. So, Rawls concludes, legitimate government activity must be restricted to the fair distribution of the fruits of social cooperation between free and equal citizens. This justifies the priority of the justice over the good.

I will not try and cover all of these issues here; instead I will try to provide a flavor of recent debates in analytical political philosophy by focusing on the following questions. I will start not with justice, but with something even

more fundamental: the justification of coercion. It is widely acknowledged that coercion is *prima facie* wrong and so impermissible, so what makes it permissible for a state not just to use but also to monopolize coercive force? The question of legitimacy raises two further questions. First, is it legitimate to use force against people who have no obligation to act in the ways that state commands? This is the question of political obligation. Second, can the use of force be legitimate if the distribution of benefits and burdens are blatantly unfair? If we think that justice matters for legitimacy, then we need to ask what counts as a just distribution of benefits and burdens. This is the question of (distributive) justice. Both questions lead to a number of further questions. As to the question of political obligation: do people acquire political obligations by doing something (giving consent, or voluntarily engaging in cooperation)? Or is political obligation independent of one's will. The former theories are "voluntaristic," while the latter are "non-voluntaristic."

In the first section of this chapter I will discuss this question by focusing on the versions of the two types of theories I find most compelling: consent theory and the theory of natural duty of justice. The first claims that political obligation can only be grounded on people consenting to be ruled by the state. The second section claims that there is a natural duty to uphold and promote just arrangements, and that obedience to the state is necessary to discharge this duty. As to the question of justice, I will focus on the issue of its site and scope. By the "site" of justice I mean the kind of entities that are regulated by principles of justice. In other words, the kind of things to which the predicate "just" can be properly applied to. On a popular view defended by Rawls, principles of justice regulate only the main institutions of society, but not the everyday behavior of people living under those institutions. G. A. Cohen (2008) challenges this view by arguing that there is no sufficient reason to restrict the site of justice to institutions alone. The "scope" of justice refers to the group of entities to which principles of justice apply to. Recent debates turn around the question whether there are sound reasons to restrict the scope of justice, or whether principles of justice apply globally. Reasons for restricting the scope include the fact of coercion, the normative significance of common nationhood, and issues of cooperation and reciprocity. Alternatively, one might defend an account of justice with a global scope. In this view, there are no compelling reasons to restrict justice to nation-states. In the third section I will defend a qualified version of global justice.

1 Obligation and Natural Duties

The vast majority of people in the world live in a state-like organized political community. By virtue of this fact they need to comply with a number of laws

and rules, which, unless followed make people liable to a variety of sanctions that go from monetary fines to incarceration. An essential feature of the state, then, is that it exercises coercive power over those under its command. Thus, a distinctive task of political philosophy is not to provide an explanation of political power, but to provide a justification for it. Because the state coerces us in many different ways, political philosophers have tried to find a general theory that legitimizes such coercion. In order to justify obedience to the state we need a theory of political obligation that establishes three theses. First, it needs to ground a *general* presumption for obedience: it needs to explain why we should obey most laws. Second, it must be *universal*: it must explain why everyone (within a political community) should obey the law. Third, the law is to be obeyed partly because it is the law and not merely because of its content. This condition of *content-independence* is needed because the law gives us a special kind of reasons. Imagine that Carl wants to cross the street. Betty suggests him not to since there is heavy traffic and he might be run over by a car. The content of Carl's deliberation should, clearly, be affected by Betty's suggestion: when deciding whether to cross the street he needs to take into account the heavy traffic. Imagine a second case in which Carl owes obedience to Ana. In this case, Ana forbids Carl from crossing the street because of heavy traffic. In this case, Carl's deliberation is affected by Ana's, but in a different way. Ana's command gives Carl a content-independent reason to act in certain way, in this case not to cross the street.¹

Philosophical anarchists respond that such justification is either not possible or highly unlikely. Wolff (1970) argues that we have a duty to be autonomous, that is, to judge the reasons for an action purely on their merits. According to him, satisfying the content-independent condition for obedience is a threat to moral autonomy, and according to his Kantian view of agency, we are duty-bound to be autonomous. He does not argue that we should disobey the law; he merely says that the fact that a command is the law does not give us an independent reason for action. His conclusion is strong: political obligation is unjustifiable. Since the publication of his work, much discussion in this area has been dedicated to refute Wolff's challenge (Raz 1986; Simmons 1979, 2001; Christiano 2008). A more modest approach is developed by skeptics like A. J. Simmons (1979, 2001). He does not argue that it is impossible to establish political obligation, but only that so far it has not been proved that state coercion is legitimate. His approach consists in taking all the popular arguments for political obligation and showing that they fail.

1.1 Consent Theories

A traditional response to the anarchist's challenge claims that political obligation is based on the consent of those ruled. Traditional social contract theorists

such as Locke and Rousseau argue that a coercive but legitimate political community is created by gathering the consent of its members. Social contract theorists start with a presumption of liberty and argue that by consenting to the authority of the state we can square obedience with freedom. This seems attractive because we tend to think that people should have an important degree of control over the constraints that apply to them. By showing that coercion has been consented to the problem raised by the anarchist is solved. The idea of consent plays the role of authorizing an agent to do something that would be otherwise wrong. For instance, although it is usually wrong to punch people in the face, if two boxers consent to fight, then they are allowed to perform this action. The consent basis of political obligation expands from these common cases: if you consent to be ruled by the state, then you acquire an obligation to do as it commands.

There are, however, several problems with the consent theory of obligation. Before we examine them, it is necessary to clarify three conditions that any act of consent needs to satisfy in order to be morally binding (Klosko 2012). First, only some agents are able to give binding consent. Only competent adults can change the moral landscape by consenting: children and people with severe mental disabilities are unable to give consent. Second, in order for consent to be valid, an agent needs to be aware of what he/she is consenting to. If a person consents to a medical treatment but he/she is misled about its details, say, he/she is not aware about the foreseeable secondary effects, his/her consent is invalid. Third, a person has to be free to give consent. Surrendering your wallet at gunpoint does not imply valid consent. One can object at this point that you are still free to keep your wallet (Steiner 2006); after all you can choose to be killed. The natural reaction to this objection holds that refusing consent has to be not unreasonably costly or difficult. Clearly this qualification requires further refinement, but it seems to be on the right track. If you could consent to everything that it is possible for you to do, then it is hard to see how consent can make a moral difference. In a nutshell, then, consent is valid when a competent agent agrees to something which he/she is aware of, and the costs of withholding consent are not too high.

Let's turn back to the problem of whether consent can ground political obligation. The most obvious difficulty for this view is that, in fact, most people have not given consent to the political institutions that rule them. Generally individuals are born into a political community and remain there till death. There is no point at which they consent to obey the state. Notice that even though some citizens have given express consent to the authority of the state (naturalized citizens are an obvious example), consent theory requires that all citizens (or at least the great majority of them) have given consent. Otherwise it fails to satisfy the condition of universality discussed in the previous section (but see Otsuka 2003).

Consent theorists have addressed this problem by distinguishing between explicit and tacit consent. They argue that although we have not explicitly consented to the authority of the state, we have tacitly agreed to be ruled by it. Instances of tacit consent are common in our lives: when we grab a bottle of milk from the supermarket aisle we tacitly agree to pay for it before we leave. When we order food from a restaurant we tacitly agree to pay the bill, etc. Tacit consent is ubiquitous and makes our lives easier. In this fashion, Locke writes “that every Man, that hath any Possession, or Enjoyment, of any part of the Dominions of any Government, doth thereby give his tacit Consent, and is far forth obliged to Obedience to the Laws of that Government” (Locke 1988, II, paras. 121, p. 348).

Tacit consent views suffer from a deep problem. It is extremely difficult to determine exactly what actions count as “consent,” and what it is, precisely, that we are consenting to. In trivial cases like the ones mentioned above there are social practices that create patterns of expectations. These practices help to explain why we have to follow the rules. More importantly, though, these practices are backed up by political institutions that determine the obligations we have. Just by entering a shop we have not consented to buy anything, but by ordering a meal in a restaurant we consent to pay for it. The indeterminacy of tacit consent comes primarily because consent is an intentional concept (Simmons 1979, 77): just as in the case of promises, one cannot give unintentional consent. The difficulty is to establish what acts count as signs of tacit consent. The impossibility of determining the conditions of consent makes tacit consent unconvincing.

A second way on which consent (both explicit and tacit) fails to generate obligations is pointed out famously by David Hume. According to the consent theory, “we may as well assert, that a man, by remaining in a vessel, freely consents to the dominion of the master; though he was carried on board while asleep, and must leap into the ocean, and perish, the moment he leaves her” (Hume 1994, 172).² The objection is best read, I believe, as pointing out the high costs of refusing consent. In the contemporary world, what options does a person have were he/she to refuse consent? Emigration or imprisonment are the only possibilities. But this is implausibly difficult or costly. On the one hand, many people are unable to emigrate even if they want to. Given the discretion that states have on immigration policy, a person might be refused entry to any country except his/her own. On the other hand, even if he/she could emigrate, the costs of doing so are too high. One has to break with one’s culture, friends and family and other ties. For this reason, consent theory fails the third condition for valid consent.

There is nevertheless a third possibility open to consent theorists. They can move from actual to hypothetical consent (Kant 1991). This alternative involves asking whether someone would give consent to the institutions that

rule them rather than looking for signs of actual (be it explicit or tacit) consent. This move seems attractive because it manages to reconcile the intuitive appeal of consent theory (the general presumption of liberty) with the difficulties explained above. In this view, political obligation is justified if persons would have given consent provided they had the opportunity. The fact that they did not is no objection. Unfortunately, this view is not unproblematic. As Dworkin writes “[a] hypothetical contract is not simply a pale form of an actual contract; it is not a contract at all” (Dworkin 1975, 18). The fact that I would have agreed to something is insufficient to put me under an obligation to do this thing. Hypothetical consent does not ground actual obligations.

Does this mean that we should reject consent theories? In my view, consent theories are still relevant to political theory, but they are useful for addressing a different question. The objections discussed here do not imply that consent plays no role in establishing moral obligations but only that it is an inadequate source of political obligation. Of course, actual consent can create obligations; if I make a promise, I have an obligation to keep it. More importantly, we can still use consent theories to understand questions of legitimacy (Waldron 1993a, 47). So, a government has good reasons to ask whether people would agree to a piece of legislation even though they do not acquire a political obligation because they would have consented to it. For instance, recall Rawls’ principle of legitimacy which holds that “our exercise of political power is fully proper when it is exercised in accordance with a constitution, the essentials of which all citizens as free and equal may be reasonably expected to endorse in the light of principles and ideals acceptable to their common human reason” (Rawls 1996, 137).

1.2 Natural Duties of Justice

Let’s return to the problem of political obligation. An alternative that has been popularized in recent years appeals to a *natural duty of justice*.³ Broadly speaking, there are two families of views on political obligation: according to *voluntaristic* theories obligation is established by an act performed by the citizen. The most popular requirement, as we have seen, is an act of consent. *Non-voluntaristic* theories, on the other hand, require no action on the part of a person. Natural duty theories are non-voluntaristic in this sense. Natural duties apply to us just in virtue of our being moral agents. These duties “apply to us without regard to our voluntary acts”; they “obtain between all as equal moral persons” (Rawls 1999a, 98–9). The clearest example of a natural duty is the duty not to harm. A moral prohibition on harming others is independent of any previous action, any kind of relation, or any form of association. It applies to me just in virtue of my being a moral person. Slightly more controversially, there is a natural duty to help others at no significant cost: if a

child is drowning in a shallow pond I have to save him/her, no matter what I (would) have agreed to, and no matter whether he/she and I stand in any kind of relationship.

Rawls expands natural duties to what he calls the “natural duty of justice.” He writes that it “requires us to support and comply with just institutions that exist and apply to us. It also constrains us to further just arrangements not yet established, at least when this can be done without too much cost” (Rawls 1999a, 99). In virtue of its non-voluntaristic nature, this account is able to avoid the problems faced by consent theories.

Nevertheless, it has also been criticized in several ways (Simmons 2005). The main objection to the natural duty view is that it fails to explain why a person should obey the laws of a particular state instead of a different (equally just, or sufficiently just) state. This is the “particularity” objection (Simmons 1979, 2005; Waldron 1993b; Wellman 2005). Why should citizens of Hungary pay taxes and obey the laws of Hungary instead of those of the Czech Republic? This objection exploits the fact that a natural duty of justice is an imperfect duty. By this it is meant that the duty is, in an important sense, underdetermined. The natural duty tells us that we must uphold just institutions and promote justice. However, it does not tell us how much of an effort we must make to promote justice, nor it tells us which institutions we should uphold. Following Rawls’ formulation, we could argue that the only institutions that “apply to us” are those of the state we live in.⁴ This reply is problematic because it is question begging: it assumes that the institutions that apply to us are those of the state we reside in. The question is, then, in what sense *these* institutions apply to us. Before moving on to answer this question, notice that according to its proponents the duty of justice commits us to respect and not interfere with the institutions of other states (Waldron 1993b). So, in a sense, the fact that there are some just institutions constrains our actions even if those institutions do not apply to us in the same manner that the institutions of our state do.

In order to address the particularity problem we need to establish what the role of political institutions is. I will argue that political institutions have three basic functions, and that the particularity problem can be (partly) solved by getting these functions right. First, in any large-scale cooperative venture there is the need to coordinate action. Many social goods are the product of cooperation and, without an agent that coordinates behavior we would be unable to secure these goods (but see Lewis 1969 and Olstom 1991). Without an agent that establishes whether we drive on the right or on the left, moving around would be very much ineffective. Moreover, we also need to solve assurance problems. It is well known that in cooperative ventures assurance is essential (Rousseau 1993, p. 87). It would be irrational of me to keep performing a collective act if I suspect that no one else is doing it too. Social cooperation also

provides public goods. A defining property of public goods is that they are non-excludable: if they are produced, it is impossible to exclude anyone from their enjoyment. Clean air, security, the rule of law are all examples of public goods. However, because of their non-excludability public goods are susceptible to free-riding. If free-riding is widespread, then the availability of public goods is threatened. In this case too we need an agent who can eliminate it (Klosko 2005).

Second, political institutions are required to adjudicate claims in the case of conflict. Participants might disagree (in good faith) about how to distribute the fruits of social cooperation. In order to solve these controversies, we need to establish a higher court of appeal that settles them. The need is even more urgent when we consider that not everyone will act in good faith. Some people will violate the rights of others. An enforcement agent that punishes rights' violators is necessary to secure social cooperation.

Third, the natural duty of justice faces additional problems of indeterminacy. Even if it tells us *what* principles of justice should be implemented, it will not always be sufficiently clear *how* to implement them. Take for instance Rawls' "difference principle."⁵ It is impossible for individuals to decide by themselves what actions they should perform in order to conform to this principle. The principle could be satisfied by a variety of schemes; different levels of income tax, VAT, inheritance tax, etc. What matters for the successful implementation of the principle is that everyone cooperates in a single scheme. The fact that there are number of different ways in which justice can be realized supports the claim that the state establishes justice; in "a particular community, the state determines what justice requires in the relations between individuals" (Christiano 2005). Notice, however, that this claim is compatible with the assumption that there are principles of justice that are independent from the existence of certain institutions (ibid.). To sum up, we need a state in order to establish justice by coordinating collective action, reducing (eliminating) free riding, and punishing offenders.

Now we can address the particularity requirement. In order to discharge our natural duty of justice we need to start solving the collective action problem with our neighbors. Since we engage in cooperation with individuals close to us, we have an obligation to establish justice with them first. Otherwise we would fail to secure any cooperation. It would be irrational to wait until we have an institution that solves all possible conflicts in the world before establishing local institutions (Waldron 1993b). The division of states we have now addresses this problem (to a certain extent). This explains why Hungarians have to comply with Hungary's institutions and not with Czech institutions. It would be impossible for Hungarians to sort out the cooperative dilemmas they face if they were complying with institutions in the Czech Republic. There are two things to be noticed about this argument. First, it acknowledges

that the state system we have now is contingent. We could have different state boundaries, and they would be appropriate insofar as they were effective in solving the problems described above. Moreover, the particularity challenge is partly solved, since I have only given instrumental reasons for supporting the institutions of our state: imagine a system in which Hungarians paid taxes to the Czech Republic, and Czechs paid taxes to Hungary. Assuming that the rates of taxation were fair, there would be no objection to this. Under this arrangement, some Czech institutions would apply to Hungarian residents, and some Hungarian institutions would apply to those in the Czech Republic. Other things being equal, we could then discharge our duty of justice in this way too.

An advantage of the natural duty account of political obligation is that it helps us to understand why the state that applies to us is authoritative, and it also helps us to see why we generally have a content-independent reason to obey the law. According to Raz, the “normal way to establish that a person has authority over another person involves showing that the alleged subject is likely better to comply with reasons which apply to him [. . .] if he accepts the directives of the alleged authority as authoritatively binding and tries to follow them, rather than by trying to follow the reasons that apply to him directly” (Raz 1986, 53). The state is authoritative because by establishing justice, we are more likely to discharge our natural duties than by following our own judgment. Authoritative commands are exclusionary reasons in the sense that they cancel the reason we have for judging a case on its merit. This means that we have a reason to comply with the law because it is law. For instance, imagine that the income tax rate is 30 percent. According to the difference principle, we have a duty to make the worse off as better off as possible. The fact that the law says that we have to pay 30 percent income tax gives us a reason to exclude judging whether a 30 percent rate is an adequate way to realize the principle. But the law is only authoritative because we are more likely to realize the difference principle by following the law than by making our own judgment.⁶ This is so, because the state establishes justice and solves cooperation problems. If we judged whether maximizing the prospects of the worse off requires a tax rate of 20 percent instead of 30 percent, we would end up failing to discharge our duties of justice. Now, this clearly does not mean that states that are not perfectly just are illegitimate, but it suggests that some level of justice is a necessary condition for legitimacy. In addition to justice, we might require that the state is democratic, or that its policies are justified in certain ways (Rawls 1996; Christiano 2005).

The arguments defended so far clarify why the state has authority, and why we have an obligation to comply with its commands. The argument puts justice at the center of political theory. The argument is, however, unclear about the “site” of justice: it does not tell us to what kind of entities “justice” is a

predicate of. One might think that principles of justice apply to political institutions but not to the activities of people living under them. On this view, institutions constrain the range of actions a person can perform, but there is no need for a person to apply the principle of justice to her daily life. He/she has, so to speak, delegated duties of justice to the state. This argument has been subjected to a powerful critique by Cohen, to which I now turn.

2 The Site of Justice

2.1 Cohen's Critique of Rawls

There are social factors that often lead to individuals enjoying a privileged position in society. Liberal egalitarians argue that these privileges are often unjust. So, for instance, they tend to condemn a system in which people's prospects of life are (partly) determined by their social class: cast or aristocratic systems are condemned because they allow morally arbitrary features to play an important role in individuals' life chances. Rawls also extends this assumption to natural facts (but see Clayton 2001). According to him, the possession of marketable natural talents is as arbitrary as one's race, sex, or social class. He writes that "[w]e do not deserve our place in the distribution of native endowments, anymore than we deserve our initial starting place in society" (Rawls 1999a, 89).⁷ This observation gives Rawls' theory a strong egalitarian bias. Since individuals do not deserve their talents or advantageous social position, inequalities have to be justified without appealing to the concept of "desert." As we have seen, Rawls' difference principle provides one such justification.⁸ The idea is that an inequality is justified because in its absence the worst off would be even more disadvantaged. The standard use of this argument is related to giving economic incentives. The view is that those who are more productive will increase their productivity even more if they are offered incentives, and that the least advantaged will benefit by the increase of society's total output. To illustrate, imagine an arrangement in which the better off earn 25,000 and the worse off earn 15,000.⁹ The better off could demand a reduction of top tax arguing that it is necessary to benefit the worst off. Under the new arrangement, the worst off will earn 20,000 and the better off will earn 40,000. Rawls concedes that inequality will increase but, insofar as it raises the level of the worst off, the tax cut is just. The incentives argument is reconstructed by Cohen as follows:

P1 There is a normative premise: "Inequalities are just if and only if they are necessary to make the worst off people in society better off than they would otherwise be" (Cohen 2008, 119). This is a reformulation of the difference principle.

P2 There is a factual premise: Unless top tax goes down, the rich will not work as hard as they could, failing to improve the situation of the worse off

C) The top tax should be lowered.

It is unclear what exactly “necessary” means in the first premise. Cohen offers two readings of the principle: the strict principle “counts inequalities as necessary only when they are, strictly, necessary, necessary, that is, apart from people’s chosen intentions”; the lax principle “countenances intention-relative necessities as well” (ibid., 69). Cohen argues that a society in which people affirm the difference principle as a principle of justice must be committed to the strict reading.¹⁰ Imagine that we ask if the factual premise is “true because the rich are unable to work at [a higher tax rate] as hard as they do at [a lower tax rate], or it is because they are unwilling to work that hard at [a higher tax rate]” (ibid., 48). In normal circumstances, the rich will need to acknowledge that incentives are required because of their unwillingness to work harder at a higher tax rate.¹¹ We can understand this as a dilemma: either the rich endorse the difference principle, in which case incentives are unjust, or the rich do not endorse the principle, in which society is also unjust. Let’s start with the first horn. The factual premise establishes that incentives are necessary for advancing the position of the worse off, but incentives are necessary only because the rich demand them. They could work harder without the extra pay, but they decide not to. For this reason, the incentives are unnecessary and unjust.¹² Moving to the second horn, the rich could insist on getting the incentive and recognize that it is unnecessary (in an intention-insensitive way). This move, however, is unsatisfactory: according to Rawls a “well-ordered” society displays “full compliance,” that is, everyone accepts the principles of justice, and this is common knowledge. If the rich recognize that they do not endorse the principle of justice society fall short from full-compliance, and to that extent be unjust. So, Cohen concludes that the only plausible principle is the strict one. And, moreover, he argues that in a just society individuals will refrain from demanding salaries that create unnecessary inequalities. This means that principles of justice must not only guide the design of institutions (tax law, for instance), but also the “structure of response lodged in the motivations that norm everyday life” (ibid., 123). Cohen refers to these responses as an *ethos* of justice.

2.2 The “basic structure” Objection

Rawlsians can defend the argument for incentive inequality by appealing to the “basic structure objection” (Cohen 2008, 116–51). There exists, they might

argue, a division of labor between major social institutions and individuals' private actions. The basic structure is "the way in which the major social institutions distribute fundamental rights and duties and determine the division of advantages from social cooperation" (Rawls 1999a, 6). Defending this division of labor, Rawls writes that

the principles of justice [. . .] apply to the main public principles and policies that regulate social and economic inequalities. [. . .] The difference principle applies [. . .] to the announced system of public law and status and not to particular transactions or distributions, nor to the decisions of individuals and associations, but rather to the institutional background against which these transactions takes place. (Rawls 1996, 282–3)

So, according to the Rawlsian defence, Cohen misunderstands the "site" of justice: principles of social justice apply to institutions rather than to everyday life. Justice does not require an egalitarian ethos. Cohen attacks this reply by noting an alleged ambiguity in Rawls' formulation. It is unclear *what* institutions belong to the basic structure and *why* they belong to it. In the previous quotation (and many others that Cohen notes) Rawls seems committed to a coercive reading of the basic structure: for an institution to be subject to the principles of justice it is necessary that it is coercive.¹³ This reply seems natural, since as we discussed in the first section, coercion always needs justification. One could reply, however, that if coercion is morally significant, it should be so because of the profound effects it has, or could have, (Pettit 2008) on people's lives. Rawls seems to acknowledge this when he writes that the "basic structure is the primary subject of justice because its effects are so profound and present from the start" (Rawls 1999a, 7).¹⁴ This answers the "why" question. But this response, Cohen argues, is fatal for Rawls' view on the "what" question. For Rawls faces a dilemma. If he sticks to a legally coercive account of the basic structure, his view seems arbitrary. There seems to be no non-arbitrary reasons to think that only these institutions belong to the basic structure. As we have argued, the normative significance of the basic structure is that it has profound effects in people's lives (but see Williams 1998). Although it is true that legally coercive institutions produce these effects, non-coercive institutions produce them too. Rawls could opt for a broader definition and include non-coercive institutions that depend on conventions, traditions and informal social rules.¹⁵ On this definition, however, the distinction between institutions and behavior that occurs within institutions collapses (Cohen 2008, 135).¹⁶ If Cohen's argument is sound, then it is impossible to rule out everyday behavior from the purview of justice.

2.3 Possible Replies

It is not clear to what extent Cohen is right. Although I believe that his critique of Rawls has some plausibility, Rawlsians can still offer the following considerations. First, the standard picture of a well-off person who votes for a higher progressive income tax, but that in his daily life tries to make as much money as possible, may not correspond with the kind of persons that would live in an ideal Rawlsian society. This might be so because taken people's motivations as given, there might be other social arrangements that neutralize their self-seeking behavior. Moreover, at any rate, we can speculate that the level of inequality we see in our world would be incompatible with a society governed by Rawlsian principles (Rawls 1999a, 47–101). Even though we cannot say exactly how much inequality would exist in that society, it is important to bear in mind that inequalities are constrained by what Rawls calls the "fair value of political liberties" (ibid., 194–200) and the principle of "fair equality of opportunity" (ibid., 73–8). Moreover, Rawls preferred economic system, "property-owning democracy" seems more egalitarian than any welfare state capitalism (see the collection of papers in O'Neill and Williamson 2012). Rawlsians, then, can concede that it is theoretically possible that a just society will tolerate large inequalities, but will insist that in practice this is highly unlikely. This point is not philosophical, of course, but it is relevant because Cohen seems to offer our society as an example of incentive-generated inequality.

Second, recall the natural duty of justice discussed earlier. Individuals have a natural duty to uphold and promote just institutions. Because of this duty, it is not the case that principles of justice make no demands from individuals. Cohen himself agrees that an egalitarian ethos would allow individuals some space to benefit themselves. Through the introduction to an "agent-centred prerogative" (Cohen 2008, 61, 387–94) individuals might pursue, to a limited extent, their own projects in spite of creating some inequalities.¹⁷ Although the size of the prerogative is vague, it would not tolerate the kinds of inequalities that the lax difference principle permits. Still, it might tolerate more inequality than one might think originally. As Estlund has argued, the realization of other important social values might require an important degree of discretion in using one's productive capacities. For instance, fraternity might justify asking for higher salaries in order to perform a socially beneficial job (Estlund 1998).¹⁸

Third, it is unclear, however, whether personal prerogatives are an all-things-considered permission, or whether they can be incorporated within egalitarian justice. For one could argue that justice is compromised when unnecessary inequalities exist, but that justice is not the only value that matters. We need to balance these values. Alternatively, one could argue that justice is compatible with these prerogatives. When these inequalities exist, society is no less just. This debate mainly concerns what the concept of justice is. Does

justice already incorporate values such as fairness, equality, and liberty? Or is justice a matter of equality that competes with other values like fairness and liberty? If one takes the former suggestion, Cohen's and Rawls' position might not be that far apart. Both judge that some inequalities are acceptable, but the justification for those inequalities is different. For Rawlsians it is sufficient that the inequality benefits the least advantaged, whereas Cohenians want to know the reasons that justify the inequality: for instance, they might condemn an inequality justified by greed, but accept one justified by a duty to care after one's ill mother.

Fourth, a different strategy tries to find a middle way between Rawls and Cohen. Even if principles of justice do not apply directly to individuals' everyday life, the justification of those principles might (Shiffrin 2010). According to the difference principle, the possession of marketable talents is morally arbitrary. This is one of the reasons that Rawls rules out considerations of desert or merit. The fact that a person is able to kick a football well does not entitle her (per se) to benefits unavailable to others. Advantages are justified, as we have seen, when others benefit from them. But talents are not arbitrary from all points of view: they are intimately connected to our sense of self-esteem and self-respect. For this reason, occupational choice has priority over the difference principle. Shiffrin argues that a person who accepts the difference principle as a principle of justice must believe that its justification is sound. If he/she does, then he/she cannot demand higher rewards for exercising his/her morally arbitrary talents. He/she need not be guided by the principle itself, but he/she must justify his/her daily actions on similar grounds. This approach differs from Cohen's in one crucial aspect: for him, principles of justice are fully directive action guides. Principles tell us what states of affairs we should actively seek. For this reason, Cohen's position is vulnerable to the objection that the demands it makes on us are too great, an objection to which consequentialism is also vulnerable. The appeal to personal prerogatives somehow alleviates the problem but does not look fully convincing. On the other hand, for Shiffrin, principles of justice are constraints on action. A talented person has an option to choose the less productive job, but once he/she chooses an occupation the range of options available to him/her is restricted by the justification of the difference principle. For instance, a person might benefit the worst off most by being a doctor than by being a philosopher. At this stage he/she has an option between these two occupations. But, once he/she chooses his/her job, he/she has no right to bargain for a higher salary (Kagan 1989). As he/she explains: "inequality generated by talent incentives should not occur [. . .]. But, then, work that could benefit all of us and the least well off may not be performed" (Shiffrin 2010, 130).

Finally, one might ask to what extent and exactly where Rawls and Cohen disagree. In a revealing paragraph, Cohen states that his "fundamental concern

is neither the basic structure of society, in any case, nor people's individual choices, but the pattern of benefits and burdens in society" (Cohen 2008, 126). In his conception of social justice the judgments of justice assess states of affairs. His conception of equality is a telic (or goal-oriented) conception (Parfit 2002). Rawls, on the other hand, seems to have a deontic conception of social justice. He is mainly concerned with how and when individuals have claims to the burdens and benefits of social cooperation (Scanlon 2006). On Rawls' deontic conception the claims of social justice are reasons for action. The fact that a situation is unjust implies that citizens have a reason to correct it. And, from the natural duty of justice, we have reasons to deal with these claims institutionally. Cohen is not necessarily committed to this view. An unjust state of affairs might give us only reasons to regret. Perhaps features of human nature make it impossible to achieve social justice. This fact should not change our judgments with regards to what is just. As he argues, principles are "fact-insensitive" (Cohen 2008, 229–73). Insofar as they have different conceptions of justice, at some level Rawls and Cohen talk past each other. Although it is still up for debate which is the best conception of justice, since they address different questions it may be the case that both are adequate at responding their own question.

3 The Scope of Justice

Liberal writers have often assumed that political societies are self-contained and closed: "[i]ts members enter only by birth and leave it only by death" (Rawls 1996, 12). Although this assumption is useful to understand domestic justice, it is incomplete since it leaves unanswered how we should treat justice at the global level. Once we drop this assumption, it becomes relevant to ask questions about the scope of justice. By "scope" we mean the range of application of principles of justice; the sets of people covered by these principles.

3.1 Cosmopolitanism

We can start by asking whether membership of a political community (or nationality, for short) has any moral significance. Pogge famously argues that "[n]ationality is just one further deep contingency (like genetic endowment, race, gender and social class), one more potential basis of [. . .] inequality" (Pogge 1989, 247). This seems a natural response. At first sight, nationality does not seem to track any morally relevant factors: individuals are not responsible for being born into a poor or rich country any more than being born into a rich or poor family. If we do not want family circumstances to affect someone's chances, why should we allow them when they arise from national

circumstances? If this view is correct, we can endorse a form of cosmopolitanism according to which “there are global principles of justice that include all persons in their scope” (Caney 2008, 388).¹⁹ An advantage of this position is its ecumenism: it can be endorsed by luck egalitarians²⁰ but need not be restricted to them. For instance, according to Simon Caney’s humanity-based view, “persons should be included within the scope of distributive justice simply because they are fellow human beings” (Caney 2011, 507). Note that this view is silent about the content of principles of justice: by itself it does not dictate what justice requires. It is compatible with the idea that justice requires that no one falls below a determined threshold,²¹ or that everyone faces equal opportunities (Caney 2005). It is also compatible with the idea that there is a mismatch between global and domestic principles of justice. For instance, it could be that domestic justice requires equality among co-nationals, but that global justice merely requires that everyone is secured a decent life. Explaining this asymmetry between these levels is an open, and not easy, task.

3.2 Relational Accounts

The naturalness of cosmopolitanism places the burden of proof on those who think that nationality is not morally arbitrary. Authors who reject cosmopolitanism tend to argue that principles of justice obtain in virtue of there being a special relationship between individuals. This family of views are usually called “relational” (Sangiovanni 2007), or “associational” (Caney 2011). There is a simple Rawlsian argument against cosmopolitanism.²² As we discussed in the last section, Rawls thinks that principles of justice apply only to the basic structure of society. But, the argument continues, there is no global basic structure. Therefore, there are no principles of global justice. Two strategies can be used to challenge this. First, one could follow Cohen and deny the first premise. I will not repeat here the reasons that motivate this possibility. Second, one could accept the associational view, but reject the second premise. For instance, Pogge (2002) argues that there is a shared institutional global order given the level of (economic and political) interaction between states. This fact indicates that principles of justice apply globally. Institutions such as the World Trade Organization (WTO), the International Monetary Fund (IMF), the different conventions and terms of international trade, all of them have a deep impact on people’s life. Global warming and climate change only exacerbate the problem even more.

The global order has (at least) three noteworthy features. First, it is causally responsible for creating and maintaining world poverty. It is impossible to fully track poverty to natural causes, such as disease or lack of natural resources. It is also impossible to explain poverty arising from only to domestic factors. Even if domestic governance in poor countries is deficient, the fact

that governments are recognized globally, have borrowing rights and are engaged in international trade (selling natural resources and sometimes buying arms) makes the global order responsible (if not uniquely so) for poverty (Pogge 2002, 200–1). Second, it operates by violating the poor's negative rights.²³ Pogge's argument goes beyond the idea that alleviating poverty is a matter of providing help to the needy. He argues that individuals are actually harmed by the global order. Even if some people think that positive duties do not exist, the existence of negative duties is widely recognized. Third, the global order is imposed by the rich on the poor (Pogge 2002, 199): the rules that govern international trade, the WTO, the IMF, and so forth are decided by rich countries given the imbalance of power in negotiations, and military and economic strength (Pogge 2002, 200).

One could challenge Pogge's conclusions arguing that the global order is not imposed on the poor. Even though one could agree that the rules are not fully egalitarian, they are not unfair since the poor (or their representatives) agreed voluntarily to them.²⁴ poor countries enter into international trade voluntarily; they negotiate and agree on the terms offered by institutions such as the WTO and the IMF. On this view the global order lacks the crucial element of coercion.

There are two possible responses to this objection. First, recall that one of the conditions for binding consent is that one is free to withhold consent, and that refusing consent is neither too costly nor difficult. Refusing to take part in global trade and abide by the conventions and rules of the WTO are incredibly costly (Walton 2009, 224). These considerations cast doubt on the voluntariness of the global order. Moreover, restrictions on immigration coerce foreigners, and are usually not justified to them (Abizadeh 2007).

Second, and more fundamentally, Nagel (2005) argues that principles of justice are created when coercive institutions are *jointly authored* by citizens.²⁵ Nagel appeals to the idea that coercion is always in need of justification. He argues, moreover, that only when "we are both putative authors of the coercively imposed system, and subjects to its norms' egalitarian principles of justice are applicable" (Nagel 2005, 128).

Nagel's argument can be criticized along the following lines. First, the idea of joint authorship is ambiguous. It may be read in a broad or in a strict sense. Imagine a colonial regime that aims to give differential treatment to the colonies. Under the strict reading the colonizers could refuse to make rules in the name of the colonized in order to avoid raising issues of justice. Because duties of justice exist only when people are joint authors, one can avoid committing injustice by refusing to rule in everyone's name. But this conclusion is embarrassing (Abizadeh 2007). In order to avoid it, Nagel concedes that we need "a broad interpretation of what it is for a society to be governed in the name of its citizens" (Nagel 2005, 129, n. 14). This concession comes at a

price, because it is not obvious that, under the broad interpretation, the kind of coercion imposed by the WTO and other international agents is not also jointly authored. Even if no one is asked about the decisions they take, in a sense these institutions are supposed to represent everyone (or make decisions in everyone's name). Otherwise, who or what exactly do they represent (Armstrong 2012, 99–100)?

Second, although Nagel has shown that coercion is sufficient for claims of justice to arise, he has given no reason to think it is necessary (Caney 2011). What reasons do citizens have to accept egalitarian duties when they are coerced? As we explained in the first section, a normal way to establish that an agent has authority consists in showing that a person is more likely to comply with the duties that apply to him/her by following the authority's commands than by making his/her own judgment. In this view, if egalitarian principles are true (and if an egalitarian state is authoritative), there must be reasons in favor of egalitarianism that apply independently of the authoritative agent and not as a consequence of it. The fact that coercion requires justification might mean only that a state has an obligation to explain the reasons for its policies (Caney 2008). It is not obviously true that justification establishes the content of what is justified. Imagine a person who rejects egalitarianism. He/she thinks that libertarian principles are the only thing that justice requires. Why should he/she accept egalitarian principles? Whatever reasons he/she has, must come from an account of pre-political morality.²⁶

This objection brings us back to cosmopolitanism. The account I gave above states that there are principles of justice that apply to everyone. This view is incomplete: it only covers the entitlements that people have, but it is silent about the duty-bearers. It does not yet say who should provide the goods necessary for satisfying individuals' entitlements. Recall that in order to discharge our natural duty of justice we need institutions. They are necessary in order to specify how we ought to discharge it. There might be a variety of institutional schemes that helps us better to comply with these duties. This task is of utmost theoretical and practical importance.

4 Conclusion

In this chapter I gave an overview of three areas of political philosophy. I argued that the most plausible theory of political obligation is one based on natural duties. I also argued that the content and scope of those duties do not require the existence of any institutional schemes, even though its implementation and full specification does. Finally I also argued that even if principles of justice need not apply to everyday choices, the justification of those principles puts important constraints on how we live our daily lives. Although the

overview is brief, I hope to have shown the importance, rigor, and sophistication of recent analytical political philosophy.

Notes

- 1 Whether the content-independent reasons that law give us are decisive is something that we can doubt: a law that requires that we whistle for ten seconds every day at 10.50 am gives us some reason to do it, but not, I think, a decisive one.
- 2 The consent theorist can rebut Hume's point by arguing that he has a right not to be carried aboard while asleep. This violation of his rights makes his act of tacit consent invalid. Hume, could, modify his example arguing that the case is the same had the person been born on board.
- 3 Other alternatives include the fair play account (Klosko 2005) and the associative duties account (Horton 2010).
- 4 Here I will assume that one's obligation is tied to residence rather than citizenship (see Wellman 2005, 46–52).
- 5 "Social and economic inequalities are to be arranged so that they are [. . .] to the greatest expected benefit to the least advantaged . . ." (Rawls 1999a, 72).
- 6 This explains why a law requiring that we hop on one foot at 3 pm is not authoritative; we have no independent reason to do it.
- 7 "The ability to make efforts and work hard is also morally arbitrary, since these dispositions depend on having a fortunate family and good social circumstances" (Rawls 1999a, 89).
- 8 Notice that the difference principle is constrained by a principle of equal liberties (including the fair value of political liberties) and a principle of fair equality of opportunity.
- 9 As I said above, I do not make any assumptions about the *equalisandum* of justice. Imagine also that under strict equality each group earned only 10,000. The example is adapted from Cohen 2008.
- 10 This assumption is plausible since we assume full compliance; that is, we assume that a person knows what the principles of justice (and their justification) are, and he/she knows that others know them too. He/she is also motivated by a sense of justice, and knows that everyone else is too.
- 11 There are exceptions to this case. Cohen recognizes some "personal prerogatives" (2008, 61–2) to benefit oneself (although to a much lesser extent than it is allowed by the lax reading of the difference principle. He also recognizes that some inequalities are necessary: imagine a football player who is unable to play unless he gets expensive massages. The strict difference principle would allow such inequalities.
- 12 They might be cases in which a person is unable, with all the will in the world, to perform his job without extra resources. Imagine a footballer who requires expensive ice-baths in order to reduce pain and play at his top level. According to both Cohen and Rawls this inequality would be justified.
- 13 Rawls seems to equate "coercive institutions" with "legally coercive institutions." In addition, one could also ask whether coercion is sufficient. Rawls is ambiguous about it, since he seems to suggest that only "major" coercive institutions are part of it.
- 14 He gives two additional reasons for the primacy of the basic structure. First, the basic structure "shapes the wants and aspirations that its citizens come to have" (Rawls 1999a, 229). Second, it secures the background conditions necessary to make

- individuals' economic transaction fair (Scheffler 2010, 132–3). Although these reasons are important, Cohen's argument can successfully deal with them.
- 15 Although it is clear that these institutions are not legally coercive, one could argue that they are coercive. This is plausible if one thinks that behavior that informal sanctions follow from nonconformity. Mill famously argues for a similar point in *On liberty*.
 - 16 The conceptual distinction is possible, but in practice both kinds of behavior are coextensional.
 - 17 Cohen also argues that burdensome and unpleasant jobs might require higher salaries. However, they might not count as creating inequalities according to his preferred notion of advantage (Cohen 2011).
 - 18 Imagine a couple who have satisfying jobs. However, one of them would like to attend design school and the other would like to be a doctor in poor areas. The only way that they can both satisfy this preference is if the latter doubles his/her salary. It seems that solidarity and fraternity permit him/her to bargain for higher rewards that are not strictly necessary.
 - 19 There are different understandings of cosmopolitanism (Scheffler 2001). In this chapter I will use the word to refer only to the definition given in the text.
 - 20 Luck egalitarianism holds that other things equal "it is bad — unjust and unfair — for some to be worse off than others through no fault [or choice] of their own" (Temkin 1993, 12).
 - 21 A very weak form of cosmopolitanism is not distinctive since very few would disagree with it (Miller 2007). This is why most cosmopolitans either appeal to global egalitarianism, global prioritarianism or a demanding global sufficientarianism.
 - 22 This is not Rawls' argument. Rawls' theory of global justice is not individualistic. He lays the foundation of justice between "peoples" not individuals (Rawls 1999b). I will not deal with his theory in this chapter.
 - 23 Negative rights tend to be understood as claims against interference. Positive rights, by contrast, involve a claim to be provided to some good. For instance, a right not to be stabbed is a negative right, whereas a right to education is a positive right. The right-holder is entitled to be provided with education. Negative rights are far less controversial than positive rights.
 - 24 The doctrine of "*volenti non fit injuria*" comes to mind. It is not clear that the argument applies to non-democracies. For it to have force it must assume that those in power have indeed the right to rule.
 - 25 He does recognize though that there are humanitarian duties owed to needy foreigners. However, these duties are neither duties of justice, nor egalitarian.
 - 26 I am grateful to JánosKis for suggesting this line of thought.

Bibliography

- Abizadeh, A., 2007. "Cooperation, Pervasive Impact, and Coercion: On the Scope (not Site) of Distributive Justice." *Philosophy & Public Affairs*, 35, pp. 318–58.
- Anderson, E., 1999. "What Is the Point of Equality?" *Ethics* 109, pp. 287–337.
- Armstrong, C., 2012. *Global Distributive Justice*. Cambridge: CUP.
- Arneson, R., 1989. "Equality and Equal Opportunity for Welfare." *Philosophical Studies*, 56, pp. 77–93.
- Caney, S., 2011. "Humanity, Associations, and Global Justice." *The Monist*, 94, pp. 506–34.

- , 2008. "Cosmopolitanism and Justice." In T. Christiano and J. Christman, eds, *Contemporary Debates in Political Philosophy*. Oxford: Blackwell, pp. 387–407.
- , 2005. *Justice Beyond Borders*. Oxford: OUP.
- Clayton, M., 2001. "Rawls and Natural Aristocracy." *Croatian Journal of Philosophy*, 1, pp. 239–59.
- Clayton, M. and Williams, A., eds, 2002. *The Ideal of Equality*. Basingstoke: Palgrave.
- Christiano, T., 2008. "Authority." *The Stanford Encyclopedia of Philosophy* (Spring edn). E. N. Zalta, ed. Available at: <http://plato.stanford.edu/archives/spr2012/entries/authority/>
- , 2005. *The Constitution of Equality*. Oxford: OUP.
- Cohen, G. A., 2011. *On the Currency of Egalitarian Justice*. Oxford: OUP.
- , 2008. *Rescuing Justice and Equality*. Cambridge: Harvard University Press.
- Dagger, R., 2010. "Political Obligation." *The Stanford Encyclopedia of Philosophy* (Summer edn). E. N. Zalta, ed. Available at: <http://plato.stanford.edu/archives/sum2010/entries/political-obligation/>
- Dworkin, R., 2002. *Sovereign Virtue*. Cambridge: Harvard University Press.
- , 1975. "The Original Position." In N. Daniels, ed., *Reading Rawls*. New York: Basic Books, pp. 16–52.
- Estlund, D., 1998. "Liberalism, Equality, and Fraternity in Cohen's Critique of Rawls." *Journal of Political Philosophy*, 6, pp. 99–112.
- Horton, J., 2010. *Political Obligation*. New York: Palgrave.
- Hume, D., 1994. *Political Writings*. Edited by S. Warner and D. Livingston. Indianapolis: Hackett Books.
- Hurley, S., 2003. *Justice, Luck, and Knowledge*. Cambridge: Harvard University Press.
- Kagan, S., 1989. *The Limits of Morality*. Oxford: Clarendon.
- Kant, I., 1991. *The Metaphysics of Morals*. Cambridge: CUP.
- Klosko, G., 2012. "The Moral Obligation to Obey the Law." In A. Marmor, ed., *The Routledge Companion to Philosophy of Law*. New York: Routledge, pp. 511–26.
- , 2005. *Political Obligations*. Oxford: OUP.
- Lewis, D., 1969. *Convention*. Cambridge: Harvard University Press.
- Locke, J., 1988. *Two Treatises of Government*. Edited by P. Laslett. Cambridge: CUP.
- Mill, J. S., 1991. *On Liberty and Other Essays*. Oxford: OUP.
- Miller, D., 2007. *National Responsibility and Global Justice*. Oxford: OUP.
- Mulhall, S. and Swift, A., 2006. *Liberals and Communitarians*. Oxford: Blackwell.
- Nagel, T., 2005. "The Problem of Global Justice." *Philosophy & Public Affairs*, 33, pp. 113–47.
- O'Neill, M. and Williamson, T., 2012. *Property-owning Democracy*. Oxford: Wiley-Blackwell.
- Olstrom, E., 1991. *Governing the Commons*. Cambridge: CUP.
- Otsuka, M., 2003. *Libertarianism without Inequality*. Oxford: OUP.
- Parfit, D., "Equality or Priority." In M. Clayton and A. Williams, eds, *The Ideal of Equality*. Basingstoke: Palgrave, pp. 81–125.
- Pettit, P., 2008. "Freedom and Probability: A Comment on Goodin and Jackson." *Philosophy & Public Affairs*, 36, pp. 206–20.
- Pogge, T., 2002. *World Poverty and Human Rights*. Cambridge: Polity.
- , 1989. *Realizing Rawls*. Ithaca: Cornell University Press.
- Quong, J., 2011. *Liberalism without Perfection*. Oxford: OUP.

- Rawls, J., 1999b. *The Law of Peoples with "The Idea of Public Reason Revisited."* Cambridge: Harvard University Press.
- , 1999a. *A Theory of Justice*. Cambridge: Harvard University Press.
- , 1996. *Political Liberalism*. New York: Columbia University Press.
- Raz, J., 1986. *The Morality of Freedom*. Oxford: Clarendon.
- Rousseau, J. J., 1993. *The Social Contract and Discourses* (ed. J. Dent) (London: Everyman Books).
- Sangiovanni A., 2007. "Global Justice, Reciprocity, and the State." *Philosophy & Public Affairs*, 35, pp. 3–39.
- Scanlon, T., 2006. "Justice, Responsibility and the Demands of Equality." In C. Synowich ed., *The Egalitarian Conscience*. Oxford: OUP, pp. 70–87.
- Scheffler, S., 2010. *Equality ad Tradition*. Oxford: OUP.
- , 2001. *Boundaries and Allegiances*. Oxford: OUP.
- Sen, A., 2000. "Equality of What?" In S. Darwall, *Equal Freedom*. Ann Arbor: University of Michigan Press, 307–30.
- Shiffrin, S., 2010. "Incentives, Motives, and Talents." *Philosophy & Public Affairs*, 38, pp. 111–42.
- Simmons A. J., 2005. "The Duty to Obey and Our Natural Moral Duties." In C. Wellman and A. J. Simmons, eds, *Is There a Duty to Obey the Law?* Cambridge: CUP, pp. 93–196.
- , 2001. *Justification and Legitimacy*. Cambridge: CUP.
- , 1979. *Moral Principles and Political Obligations*. Princeton: Princeton University Press.
- Steiner, H., 2006. "Individual Liberty." In D. Miller, ed., *The Liberty Reader*. Edinburgh: Edinburgh University Press, pp. 123–40.
- Temkin, L., 1993. *Inequality*. Oxford: OUP.
- Waldron, J., 1993b. "Special Ties and Natural Duties." *Philosophy & Public Affairs*, 22, pp. 3–30.
- , 1993a. *Liberal Rights*. Cambridge: CUP.
- Walton, A., 2009. "Justice, Authority, and the World Order." *Journal of Global Ethics*, 5, pp. 215–30.
- Wellman, C., 2005. "Samaritanism and the Duty to Obey the Law." In C. Wellman and A. J. Simmons, eds, *Is There a Duty to Obey the Law?* Cambridge: CUP, pp. 3–92.
- Williams A., 1998. "Incentives, Inequality, and Publicity." *Philosophy & Public Affairs*, 27, pp. 225–47.
- Wolff, R. P., 1970. *In Defense of Anarchism*. New York: Harper Books.

Part III

New Directions in Analytic Philosophy

Barry Dainton and Howard Robinson

When dealing with a philosophical tradition that is barely a century old, discerning the truly significant figures and movements of the past is obviously a risky affair. Contributions that appear seminal now could easily seem otherwise with the changes in perspective and perspicuity that a century or two will bring—which is why we ended our historical survey in the 1970s. Before signing off we will, however, pause to draw readers' attentions to some noteworthy recent trends and developments, and venture some tentative guesses as to issues and approaches that may figure prominently over the next decade or so. We also will be returning to the question of whether or to what extent analytic philosophy is suffering a crisis, of one kind or another.

1 Specialization and Science

We can start with a prediction that seems safer than most. Scott Soames calls the epilogue of his survey of analytic philosophy from 1900 to 1975 "The Era of Specialization," and this is why:

In my opinion, philosophy has changed substantially in the last thirty years or so. Gone are the days of large, central figures, whose work is accessible and relevant to, as well as read by, nearly all analytic philosophers. Philosophy has become a highly organized discipline, done by specialists primarily for other specialists. The number of philosophers has exploded, the volume of publications has swelled, and the subfields

of serious philosophical investigation have multiplied. Not only is the broad field of philosophy today far too vast to be embraced by one mind, something similar is true of many highly specialized subfields. (2003b, vol. II, p. 463)

That there is much truth in what Soames says is undeniable. The number of philosophy journals has increased significantly—though as part of a more general trend: academic journals as a whole have been doubling in number every 20 years for at least a century (Mabe 2003). A brief exploration of the online guide to mainstream (mainly analytical) philosophical research *Philpapers* will quickly confirm Soames' observation that there are many subfields—particularly in growth areas such as the philosophy of mind—in which large quantities of work is being done, too large for a single mind to encompass.¹ A willingness to exploit technical developments in logic and semantics has always been a characteristic of analytic philosophy, and Soames rightly draws out attention to two fields in which a great amount of high-quality but (often) highly technical work has been done in recent decades: the topic of vagueness and semantic approaches to truth. There is every reason to think these trends will continue. Analytic philosophy will, we expect, continue to generate increasingly specialized subfields; in some areas the amount of technical work, accessible fully only to specialists in those subfields, will increase. As Timothy Williamson observes, since this technical work (at its best) leads to indisputable advances, it is not to be regretted, far from it:

In many areas of philosophy, we know much more in 2004 than was known in 1964; much more was known in 1964 than in 1924; much more was known in 1924 than was known in 1884 . . . Much of the knowledge is fairly specific in content. For example, we know far more about possibility and necessity than was known before the development of modern modal logic and associated work in philosophy. It is widely known in 2004 and was not widely known in 1964 that contingency is not equivalent to *a posteriority*, and that claims of contingent or temporary identity involve the rejection of standard logical laws. The principle that every truth is possibly necessary can now be shown to entail that every truth is necessary by a chain of elementary inferences in a perspicuous notation unavailable to Hegel. We know much about the costs and benefits of analysing possibility and necessity in terms of possible worlds, even if we do not yet know whether such an analysis is correct. (2007, pp. 280–1)

Logic and semantics aside, it is likely that the long tradition of a close relationship between analytic philosophy and science will be maintained. It is not necessary to subscribe to Quine's view that philosophy's *only* proper role is to

contribute to the clarification of science to believe that philosophy can benefit from, and on occasion contribute to, empirical discoveries and developments. In the 1920s Schlick, Carnap, and Reichenbach all contributed to the assimilation and interpretation of Einstein's relativity theories, and this engagement with contemporary physics and cosmology has continued. In *The Shape of Space* (1976, 2nd edn 1994) Nerlich argues that it is a mistake to suppose—as many do—that space (or spacetime) is a featureless nothing. Since spatial structures have discernible consequences, we should accept that space is a real and concrete entity; drawing on this strongly realist view, Nerlich goes on to criticize the conventionalist accounts of space and spacetime offered by positivists such as Reichenbach. Starting in 1990s, the debate on the “hole argument” has helped clarify our understanding of the implications of general relativity for spacetime substantivalism.² Earman has worked on singularities (1995) and the metaphysical implications of versions of quantum gravity that are hostile to the reality of time (2002).³ Huw Price took Stephen Hawking to task for failing adequately to account for temporal asymmetries (1989), and then produced his own influential account (1996).

The interpretation of quantum theory gives rise to a variety of deep philosophical conundrums, and philosophers have not been shy of entering these deep waters, for overviews see Redhead 1987, Albert 1992, and Barrett 2001. Indeed, in the case of the “many worlds” interpretation, they have even made innovative contributions to physics, see Saunders et al. 2010. Whereas the relationship of continental philosophy to physics has been tainted with scandal—see Sokal 2008—the relationship between physics and analytic philosophy has been mutually fruitful (also see Loewer's contribution to this volume).

This close partnership with the sciences is not confined to physics. The same applies to the philosophy of biology, and many of those currently working in the philosophy of mind are more knowledgeable of empirical discoveries in psychology, cognitive science, and neuroscience than was routinely the case in the past, and it is generally—though not universally—assumed that philosophical investigations can be informed by empirical results. At the start of his *The Emotional Construction of Morals*, Jesse Prinz outlines his methodology:

My most obvious commitment is to methodological naturalism, because I will draw on empirical findings throughout, including findings from neuroscience, psychology, psychiatry, anthropology, cultural history and ethology. I think enduring philosophical questions can be illuminated by empirical results, and, indeed, they might not endure for so long if we use the resources of science. That said, I do not reject traditional philosophical methods, such as conceptual analysis . . . I think that method often bears fruit, but sometimes introspections clash or fail to reveal the real structure of concepts. So it is helpful to find other methods to help adjudicate

between competing philosophical theories. These other methods cannot replace philosophy. Philosophy poses the problems we investigate, devises useful tools for probing concepts (such as thought experiments), and allows us to move from data to theory by systematizing results into coherent packages that can guide future research. I see philosophy as continuous with science, and believe that we should be open to using any methods available when asking questions about the nature of morality. (Prinz 2007, p. 9)

This “methodological naturalism” (though *methodological pluralism* might be more apt) has been widely adopted in contemporary analytic philosophy, and will very likely continue to be—for some further illustrations see Section 4 of James Ladyman’s chapter.⁴

2 Mind and Consciousness

Few would dispute that in recent decades the philosophy of mind has joined the philosophy of language and logic as a “core” area in analytic philosophy, with some—though by no means all—subscribing to the view that the philosophy of mind has supplanted the philosophy of language as the more foundational. We saw in (Chapter 15) that the philosophy of mind in the 1990s was marked by an increasingly widespread dissatisfaction with reductive forms of physicalism. This continues to be the case, and indeed the range of options on the matter-consciousness issue that analytic philosophers are prepared to take seriously continues to grow: there have been very recent debates on the merits of panpsychism, see Freeman 2006, Goff 2011, and Coleman 2012. However, for those interested in consciousness, the relationship between the experiential and the physical is no longer the only game in town.

William James’ characterization of consciousness as *stream*-like is widely accepted as phenomenologically accurate, but the distinctive unity and continuity that is characteristic of experiential streams has gone largely ignored, at least in the analytic tradition, but this has begun to change.⁵ Debates concerning the nature of perception have flourished in recent years. One central issue here is the nature of the perceptual process: does it, as many suppose, invariably involve some form of intervening mental representation, or is it direct? If perception is direct, the nature of the perceptual *relation* comes to the fore—and there remains the problem of providing an acceptable account of perceptual hallucinations. If perception is indirect, we are still faced with the problem of correctly characterizing the *kind* of representation that it involves. An influential proposal in this connection is the “representationalist” (or “intentionalist”) thesis that all the phenomenal—or experiential—features of

perceptual states are identical with their representational (or intentional) contents.⁶ Support for this proposal comes from a doctrine associated with G. E. Moore, that experience is “transparent,” that is, that we are aware of *what we are experiencing* (e.g. the page in front of one’s eyes), but we are not aware of any features of our experience itself. Moore’s transparency thesis has a good deal of phenomenological plausibility, but whether it can provide the representationalist with what they need is highly debatable.⁷ For a useful survey of the competing views, see Snowdon’s chapter in this volume and Robinson 1994; Gendler and Hawthorne (2006) offer a representative collection of relevant recent essays. These issues aside, there has also been a good deal of recent interest in the differences between the senses—how to characterize the different sensory modalities—see Macpherson 2011.

Not surprisingly, the newfound respectability of consciousness has also impacted upon debates relating to the nature of the self. Parfit’s important discussion of personal identity in *Reasons and Persons* has been a dominant influence on debates in this area since its appearance in 1984, but consciousness did not feature prominently in it. Relying on an imaginative range of thought experiments, featuring teletransportation, divided brains, and rapid physical and mental transformations, Parfit argues that personal *identity* as such is of no importance; what matters—the relationship that it is rational for us to care about—is psychological continuity (i.e. overlapping chains of memories, beliefs, intentions, and the like), a relationship that normally goes hand in hand with identity, but that can in principle come apart from it. More recently, a number of philosophers have argued that selfhood and consciousness are intimately intertwined in a way that the more orthodox psychological approach adopted by Parfit fails explicitly to acknowledge—see Valberg 2007, Dainton 2008, Strawson 2009, and Johnston 2010. But although there is a consensus here that the self should be construed as a “subject of experience,” and that the latter need not be construed as immaterial Cartesian substances, when it comes to the *correct* metaphysical account of these subjects, there is as yet nothing resembling agreement.⁸

One of the more intriguing recent developments, and one that may have longer-term and wider implications, is the *phenomenal intentionality* [P1] program, as it has recently become known.⁹ It is a fundamental assumption in the philosophy of language that some things—in particular linguistic expressions, thoughts, and experiences—can have something else as their content. Much work bearing on this idea has been done over the past century in the theory of meaning, for example, truth-conditional semantics, possible-worlds semantics, teleofunctional semantics and causal theories of reference have all been developed in detail. What has less often been discussed, at least in the analytic tradition, is the relationship between intentionality (construed broadly, so as to include both meaning and reference)

and phenomenal consciousness. What discussion there has been on this topic has, until recently, been largely negative, as instanced by the work inspired by the later Wittgenstein. But over the past decade or so, this has begun to change: in a series of recent works, some analytic philosophers have argued that a number of important forms of intentionality are found originally in phenomenal consciousness—found there in a way that is autonomous with respect to the way in which intentional content attaches to expressions of the public language, perhaps even in such a way as to ground the intentionality of thought and language. Important contributors to this line of thinking include Loar 1987, 2003, Searle 1987, 1992, Strawson 1994, 2004, Pitt 2004, 2008, Horgan et al. 2004, and Kriegel 2007.

Despite some differences in terminology and doctrine, supporters of the PI approach would all agree with the following claims: (a) there is an important species of intentionality to be found in some (or perhaps all) phenomenally conscious states that is intrinsic to such states, and is determined solely by their phenomenal features, and (b) this mode of intentionality is fundamental, in that all other manifestations of intentionality are derivative of it. If the claims of the PI movement are correct, orthodox analytic philosophy of language over the past century has been radically in error, and the true source of meaning—namely, intentional content, present in phenomenal consciousness—has simply been ignored. The PI movement envisages returning to an older, empiricist (even Cartesian) tradition, one that enjoyed a long period of orthodoxy before the linguistic turn of the twentieth century purported to overthrow it. As Pitt (2008) points out, the PI doctrine sits uneasily with Frege's influential rejection of psychologism, and as we have seen, Frege's hostility to the doctrine was shared by Moore, Russell, and Wittgenstein. But Pitt—himself an advocate of PI—recommends a return precisely *to* psychologism. A radical move indeed, within the confines of analytical philosophy at least.

There is also the prospect of solving a range of problems that have arisen in recent philosophy of language. Phenomenal intentionality, assuming it exists and has the features typically attributed to it, provides a simple and apparently appealing solution to the various indeterminacy issues that continue to plague orthodox philosophy of language. Influential arguments expounded by Quine, Putnam, and Kripke all purport to show that facts about one's previous linguistic usage, behavioral dispositions, and physical constitution do not suffice to determine whether one means addition by "plus" (Kripke), cats or cherries when one talks of "cats" (Putnam), rabbits or collections of undetached rabbit-parts when one talks of "rabbits" (Quine).¹⁰ Advocates of PI—see Searle 1987 and Strawson 2005—maintain that these difficulties dissolve when the experiential dimension of meaning is recognized. Meaning in consciousness is fully determinate; indeed, this is the only place where it is

fully determinate. Hence we can be sure that we mean plus by “plus,” cherries by “cherries,” and rabbits by “rabbits,” because these are the meanings we experience when we use the terms in question.

So the PI approach offers solutions to some longstanding and worrying problems. But since these solutions fly in the face of analytic orthodoxy, they are more controversial by far than they would be otherwise. This debate has only just begun—for a sample of some opening salvos see Bayne and Montague 2011.

3 Analytic Metaphysics

For large parts of its (brief) history, a defining feature of large parts of analytic philosophy has been a cautious—if not downright hostile—attitude to metaphysics conducted in the traditional (pre-Kantian) style. The dismissal of metaphysics as cognitively meaningless by the logical positivists is the best-known instance of this stance, but it is by no means alone: the ordinary language movement was no friend of metaphysics either. P. F. Strawson may be renowned for his rehabilitation of metaphysics in the 1960s, but the brand of metaphysics that Strawson was prepared to look favorably upon was of the “descriptive” variety. An investigation into the most general features of our conceptual scheme is certainly a task worth pursuing, but falls a long way short of an inquiry into the ultimate nature of reality itself, and it is reality itself—not our ways of thinking about it—that was the target of orthodox metaphysicians.

In the light of this background, one of the more surprising developments in recent years has been the gradual growth of the now-flourishing school of “analytic metaphysics,” as it has become known. The topics under investigation are those of traditional metaphysics, for example, the nature of substances, properties, matter, change, identity, time, and most of those pursuing these inquiries—and several of the most promising of the younger generation of analytic philosophers fall into this category—are clearly doing so under the assumption that it is reality itself that they are investigating. They also bring to bear all the technical tools that analytic philosophers now have in their arsenals.

As for why this “metaphysical turn” has recently occurred, there are several contributing factors. Kripke rehabilitated the notions of “essence” and essential properties in the 1970s (as related by Lowe in his contribution to this volume). Chisholm’s writings on metaphysics, on various topics, began to be influential in the 1960s and 1970s. The work of David Lewis is another important factor. Lewis, originally a student of Quine, was quite possibly the most influential analytic philosopher among other analytic philosophers over

the last 25 years—an influence not curtailed by Lewis' early death in 2001. Although Lewis did important work in the philosophy of language and the philosophy of mind, he also took metaphysical questions very seriously. In addition to his work on various specific metaphysical issues—for example, the nature of properties, causation, persistence, dispositions, personal identity, time travel—Lewis also defended a highly distinctive, and highly controversial, global view of reality. According to his “modal realism,” the case for which Lewis elaborates in detail in his *On the Plurality of Worlds*, there exists an infinite number of logically possible universes, each of which is just as real as this universe. The notion that there are vast numbers of possible worlds goes back to Leibniz, but for previous philosophers these worlds were assumed to be nonactual—only one world, *this* one, is real. If Lewis is correct, this assumption is wrong, and concrete reality is far (far) bigger than had previously been thought.

With this audacious doctrine, large-scale systematic and manifestly *revisionary* metaphysics was clearly back on the philosophical agenda. And Lewis was not alone in being prepared to defend revisionary metaphysical views. In *Material Beings* Peter van Inwagen addressed the “special composition question”: in what circumstances do things “add up or compose something”? (1990, p. 31). The question seems innocuous enough, but van Inwagen argues that most of the obvious answers to it fail, which leads him to conclude that the only material objects that do exist are elementary (part-free) particles, and living organisms. Tables, chairs, planets, and stars do *not* exist. Whereas Lewis would add (vastly) to the ontology of common sense, van Inwagen wishes to subtract from it. In addition to these noteworthy individual contributions, as Zimmerman points out in his “Metaphysics after the Twentieth Century” (2004), even during the years when the anti-metaphysical fervor was at its highest—from 1930 to 1950, say—there were always analytic philosophers who were oblivious to the broader trend, and continued to work on metaphysics, for example, G. F. Stout, C. D. Broad, C. A. Campbell, H. H. Price, D. C. Williams, C. J. Ducasse, G. Dawes Hicks, William Kneale, and A. C. Ewing. Metaphysics did not suddenly reappear in the 1970s, it had never wholly gone away.

The range of issues and positions the new metaphysicians have debated is already impressive, but present purposes will be served by looking at just a couple of examples. As we have just seen, van Inwagen brought the issue of composition into prominence. His position is radically counterintuitive, but it is by no means the most revisionary of those on offer. Some analytic metaphysicians defend *mereological nihilism*,¹¹ the doctrine that no objects ever combine to form other objects. If this claim is correct, then the only objects that exist are the elementary particles—assuming that such exist. Others defend *mereological universalism*: the doctrine—favored by Lewis,

Sider, and many others—according to which each and every combination of objects constitutes another object. On this view, your big toe, the right half of the moon, and the black hole at the center of our galaxy go to constitute a material object, one that is just as much an object as ordinary tables and chairs. Let us take another example: what does the persistence of an object involve? The answer seems clear enough: an object persists from t_1 to t_2 if it exists at these times, and at all the times in between. While this is true, it is not the end of the story, for there are different metaphysical accounts of what is involved in existing through intervals of time. In Lewis' terminology, objects can persist by *enduring*, or they can persist by *perduring*. An object endures by being wholly present at a succession of different times. An object perdures by being spread through an interval of time in much the same way a spatially extended object extends through a volume of space. Just as your body has spatial parts existing at different locations (e.g. your head, your hands, your feet), a perduring object has *temporal parts* existing at different temporal locations (your body as it was yesterday, and as it is today). Hence, according to the perdurance theorist, when you see your friend across the street, you are not seeing the *whole* of your friend—contrary to what you might initially be inclined to say—but a mere part of them. There is little doubt that the endurance view is closer to common sense than the perdurance view. Perdurance theorists acknowledge as much: they believe their view is better able to solve a range of philosophical problems—often of an abstruse kind—than the more familiar alternative.

These debates may feature radical and revisionary doctrines concerning material bodies—see Johnston's contribution to this volume for some further examples—but they do not, on the face of it, resemble scientific debates. In fact, the relationship between these metaphysical debates and science is a contentious issue. For some analytic metaphysicians, the fact that the perdurance view coheres well with the 4-D view of reality that fits most naturally with Einstein's special theory of relativity is a significant plus point. But for others, the consilience with contemporary physical theory is of no concern: if physics and metaphysics come into conflict, it is up to the physicists to bring their theories into accord with the metaphysical facts. This willingness to give the deliverances of metaphysics priority over those of physics is also on display in recent debates on the nature of time itself. For decades, the analytical philosophy of time consisted almost entirely of the analysis of temporal terms and temporal concepts. No longer: analytic metaphysicians now construe the debate regarding the nature of time as a dispute relating to very different theories as to the large-scale structure of the universe. According to the "eternalist" or "block" view, the universe is a 4-D entity, and past, present, and future events are all equally real. For the proponents of the "growing block" view, the past and present are real, but the future is not, but thanks to the

passage of time—which consists of new momentary slices of reality being created moment by moment—the past is always growing, and the future shrinking. For the “presentist,” in contrast, only the momentary present is real; the past and the future are wholly unreal. Now, irrespective of their other merits and demerits, both the presentist and the growing block accounts require universe-wide “planes” of simultaneity, and this requirement does not sit easily with Einstein’s relativization of simultaneity to inertial frames of reference. As Einstein’s special theory of relativity is normally construed, there are no objective facts as to which events are *really* simultaneous with others. For the proponents of the growing block theory and presentism this does not matter: all that it shows is that there are facts (in this case pertaining to objective simultaneity) that go beyond those that current physics acknowledges. For a more detailed overview of these controversies see Dainton 2010.

Not surprisingly, the willingness of analytic metaphysicians to expend their intellectual energy on (what can easily seem) highly trivial, quasi-scholastic questions, and the willingness of some of them to disregard the consequences of our best physical theories, has not been welcomed by all. Here is Craig Callender commenting on the relationship between science and metaphysics:

... when I bend my fingers into a fist, have I thereby brought a new object into the world, a fist? In contemporary metaphysics, a question such as this is viewed as deep, interesting, and about the structure of mind-independent reality. Comparable questions in the literature are of whether a piece of paper with writing on one side by one author and another side by a different author constitutes two letters or one (Fine, 2000), whether roads that merge for a while are two roads or one, and whether rabbit-like distributions of fur and organs (etc.) at a time are rabbits or merely temporal parts of rabbits.

Other philosophers, by contrast, react in horror at the suggestion that these questions are deep and important. Instead they find them shallow. The reason is that it’s hard to imagine what feature of reality determines whether a fist is a new object or not. How would the world be different if hands arranged fist-like didn’t constitute new objects? And if there are debates, aren’t they easily solved? Call temporally extended distributions of fur and flesh in bunny shaped patterns “rabbits₁” and non-temporally extended such patterns “rabbits₂.” Use “letter₁” for letters individuated by author and “letter₂” for those individuated by paper. And so on. Now is there any residual disagreement about the non-semantic world? If fists really are new objects, then one imagines that philosophers of science bring two new objects into the world whenever they read this work. (2011, p. 38)

Callender is himself a philosopher of science who, in a broadly Quinean spirit, holds that science has proved to be our best guide to reality, and metaphysics is at its best when it works in close harmony with the sciences. Callender is by no means alone in harboring doubts about the approach adopted by some of the analytic metaphysicians. A sign of the renewed interest in the nature and proper place of metaphysics in the broader philosophical arena is the recent coining of the term “metametaphysics” as a label for precisely this inquiry. In a recent anthology devoted to the topic—*Metametaphysics* (Chalmers et al. 2009)—several contributors express worries over the more esoteric ontological obsessions of the analytic metaphysicians. Yablo argues that many of these questions have no determinate answers. For Chalmers, such questions are often merely “verbal” and so nonsubstantive, and he suggests a way of demarcating substantive from merely verbal issues. Hirsch adopts a similarly Carnap-inspired approach, arguing that the dispute between mereological universalists and their nihilist opponents comes down to their attributing different meanings to “there is.” According to Hirsch’s “quantifier variance” thesis, there are multiple concepts of existence, each associated with different quantifiers, each of which conforms to the standard patterns of inference of the existential quantifier, and each of these alternative senses of “exists” is perfectly adequate for describing the world. Consequently, the mereological universalist and nihilist are (in effect) speaking different languages, but making perfectly correct claims within the language each speaks. If so, the debate between them is obviously a trivial one.

4 Rising Self-consciousness

This controversy regarding the status of metaphysics is only a part—albeit an important part—of a broader movement. Most commentators would agree that over the past decade or so there has been a general rise in both methodological and historical “self-consciousness” among analytic philosophers.

Thanks to works such as Hylton’s *Russell, Idealism, and the Rise of Analytic Philosophy* (1990) and Griffin’s *Russell’s Idealist Apprenticeship* (1991), and more recently Candlish’s *The Russell/Bradley Dispute* (2007) we have a deeper understanding of the relationship of Russell and Moore to the idealist tradition against which they rebelled. Hacker (1996c) attempts to assess Wittgenstein’s place in the entire analytic tradition. The essays in Gaskin (2001) focus on the differing views of *grammar* of Frege, Husserl, Russell, Carnap, and Wittgenstein, those in Beaney (2007) focus on the different conceptions of *analysis* in early analytic philosophy and phenomenology. Recent scholarship has transformed the picture—readily available to the English-speaking world, at any rate—of the logical positivists in general, and Carnap and the *Aufbau* in particular.¹²

On the methodological side, the role and rationale of conceptual analysis and thought experimentation have figured prominently in recent debates. The “experimental philosophy” movement is an attempt—misguided in the eyes of some—to put an end to the (as they see it) uncritical reliance that analytic philosophers place on the “intuitions” that are evoked by thought experiments. The aim is to shed light on the reliability of these intuitive responses by conducting systematic empirical studies, using the techniques of psychology and cognitive science, into the mental processes that generate them.¹³ In *The Philosophy of Philosophy* (2007), Timothy Williamson argues that it is a mistake to think that philosophy is (or should be) a matter of linguistic or conceptual analysis: it is an investigation into reality itself. Williamson also argues that our intuitive responses to imaginary cases are sophisticated judgments, produced by a generally reliable *knowledge*-generating mechanism that has evolved through natural selection. For Williamson, the imagination is itself a guide to reality, albeit fallible.

Colin McGinn has recently defended the view that philosophy *is* (and should be) a matter of conceptual analysis—although McGinn does not think this is incompatible with analysis-revealing truths pertaining to reality itself.¹⁴ Chalmers and Jackson also argue that conceptual analysis of a certain kind is possible: the meanings of our expressions are such that a rational speaker is able, on reflection, to know *a priori* whether or not a term (or concept) applies in all epistemically possible situations.

In recent writings—(2006, 2010)—Chalmers makes it clear that his preferred “epistemic” construal of the 2-D modal framework is a crucial element in a far larger and more ambitious project: that of securing a “golden triangle” of constitutive connections between meaning, reason, and modality. Kant forged a connection between reason and modality by holding that what is knowable *a priori* is necessary; Frege connected reason and meaning by arguing that there is a component of meaning, *sense*, that exists over and above reference, which is constitutively linked to cognitive significance. In the final link of the chain, Carnap connected modality and meaning via his concept of *intension*, which connects a component of meaning to possibility and necessity. The golden age proved to be short-lived: when Kripke plausibly argued that there are necessary truths that are *a posteriori*, the connection between reason (in the form of the *a priori*) and necessity was broken. Chalmers hopes to return us to the golden age: “Two-dimensional semantics promises to restore the golden triangle . . . it promises to explicate further aspects of meaning and modality that are more closely tied to the rational domain. . . . In this way, we might once again have a grip on an aspect of meaning that is constitutively tied to reason” (2006, p. 55).

5 Crisis?

We saw in the *Preface*, there are those who believe that contemporary analytic philosophy as a whole is in a troubled condition. Hacker claimed that

the movement had lost its “distinctive profile” in the 1970s; Leiter proclaims simply that analytic philosophy—in the form its critics criticize—is simply “defunct”; whereas for Biletzki and Matar it is “beyond question” that analytic philosophy has been in a state of crisis for some time. The increasing interest in historical issues and methodology that we have been surveying latterly could certainly be interpreted as responses to an ongoing and quite possibly terminal crisis. However, we believe such an interpretation, understandable as it may be, would be a mistake.

In contemporary analytic philosophy there are indeed heated disputes about the viability of metaphysics, or this or that mode of analysis. But there is nothing new here. If our whirlwind tour through the historical story, partial and incomplete though it is, demonstrates anything beyond any question it is this: the notion that analytic philosophers were once united behind a particular conception of philosophy—of philosophical analysis—is a myth. There never was a golden age, or at least not one with this monolithic character. The hegemony of the decompositional form of conceptual analysis that Russell and Moore advocated around 1900 did not endure for long, at least in Russell’s case: by 1905, with his theory of descriptions, he had discovered the advantages of the very different logical or “transformative” analysis. Not long after he added a “constructive” mode of analysis to his arsenal. It was this form of analysis that he deployed when engaged in reducing the material universe to a logical construction out of sense-data—and to which Carnap turned during his own phenomenalist phase. The commitment to the possibility of an ideal language, one that could mirror the structure of reality, that Russell and Wittgenstein shared in their logical atomist phase was rejected by Carnap, with his framework pluralism. The various “linguistic” modes of analysis adopted by the later Wittgenstein and the ordinary language philosophers was different again, and had little in common with Quine’s naturalistic program, or the truth-conditional semantic approach of Davidson, or the possible worlds semantics of Lewis and others. The condition of *competing and co-existing* methods—a methodological pluralism—has long been the norm in analytic philosophy, and will likely continue.

As for the increasing interest in the history of the analytic movement, this is largely due to the fact that it is now old enough, after barely a century, to *have* a history. And inevitably, this history reveals that the past, even the fairly recent past, is more of a foreign country than it is often supposed to be. The views of the early pioneers turn out to be sometimes stranger, invariably more complex and dynamic—particularly in Russell’s case—than the simplistic caricatures many analytic philosophers are themselves raised on. We have tried to bring a little of this richness out in our history of the movement. It is also the case that the history of analytic philosophy lives on: the concerns, doctrines, and arguments of the past are being actively engaged with by contemporary

analytic philosophers. Dummett's *Frege: Philosophy of Language* (1973) brought Frege to center-stage in the philosophy of language.¹⁵ In 1980 Hacker and Baker published the first part of their multivolume "analytical commentary" on the *Philosophical Investigations*, and seek to demonstrate the relevance of the later Wittgenstein to contemporary concerns.¹⁶ The idea that inspired first Frege, then Russell and Whitehead, that mathematics can be reduced to logic, is by no means dead: various forms of "neo-logicism" have been developed.¹⁷ And there is no shortage of other examples of the *continuity* of concerns in analytic philosophy, but we will draw matters to a close with a brief look at a small handful that are particularly worthy of note.

We saw in Part I that the Platonic Atomism of the early Russell and Moore made concepts, and hence propositions, the building blocks of reality, and thus guaranteed the intelligibility of the universe. As McDowell put it, "There is no ontological gap between the sort of thing one can mean, or generally the sort of thing one can think, and the sort of thing that can be the case. When one thinks truly, what one thinks *is* the case" (McDowell 1994, 27). Richard Gaskin has recently attempted to furnish this doctrine with new, firmer foundations. In *The Unity of the Proposition* (2008) he defends a "linguistic idealism" according to which reality is composed of propositions, which are in principle expressible in a language, and hence in principle graspable in thought. These propositions are of the "Russellian" variety: they are composed of objects and properties, and hence not invariably abstract (or nonconcrete). That propositions *are* unified—and so distinct from a mere list of words, or an aggregation of discrete concepts—is of course true. But providing a successful account of what propositional unity consist in has proven difficult: Frege, Russell, Wittgenstein, and Davidson, to mention but a few, have all developed their own theories. Gaskin argues that all of these accounts are fatally flawed, in one way or another. His own account is bold and novel. Gaskin begins by arguing that Bradley's regress is inescapable. As Bradley appreciated, the existence of any unified whole of the form aRb , where R is a relation holding between the items a and b , confronts us with a dilemma. Either R is related to its relata, or it is not. If the latter, then $[a, R, b]$ are a mere collection of unrelated items, and so not a whole at all. If the former, then we need *new* relations to connect R with a and b , and this is only the beginning, for the same problem arises for these additional relations, and so on ad infinitum. However, rather than being an insuperable problem, for Gaskin this regress is the *solution* to the problem of the unity of the proposition:

... far from being vicious, it is exactly the generation of Bradley's regress that guarantees our ability to *say* anything at all. Bradley's regress emerges not as an embarrassment, something to be circumvented by careful legislation, but as the metaphysical ground of the unity of the proposition. ... we might say, perhaps riskily, that what stops a

proposition from being a *mere* aggregate, and the corresponding sentence from being a *mere* list, is that the proposition unfolds into an *infinite* aggregate, and the sentence into an *infinite* list. (2008, p. 343)

And if reality is itself composed of propositions of the Russellian variety, then this same regress also holds reality together: in its absence, there would not *be* the totality of interrelated objects that jointly constitute the world.¹⁸

We saw above that a dispute has arisen in metametaphysics between those who believe that some of the more esoteric ontological disputes—for example, the struggle between mereological universalists and mereological nihilists—are substantive, and those who believe they are not. The doctrine of quantifier variance, as developed by Hirsch, is one way of substantiating the charge of triviality: on this view, the universalist and nihilist are merely using the concept “exists” in different ways, and making perfectly correct claims within the language each speaks. In contrast, Sider resolutely rejects the triviality charge, and in “Ontological Realism” expresses the concern that some people are losing “their metaphysical nerve” (2009, p. 384). While he accepts that an overly naïve trust in metaphysics is inappropriate, Sider argues that the quantifier variance thesis is false: there *is* a sense of “exists” which is better than others, a sense that carves reality at the joints, which accurately captures the “logical structure of reality.” This commitment to the availability of a perfect description of reality is, he argues, at the heart of analytic philosophy:

A certain core realism is, as much as anything, the shared dogma of analytic philosophers, and rightly so. The world is out there, waiting to be discovered, it's not constituted by us—all that good stuff. Everyone agrees that this realist picture prohibits truth from being generally mind-dependent in the crudest counterfactual sense, but surely it requires more. . . . The realist picture requires the “ready-made world” that Goodman (1978) ridiculed; there must be structure that is mandatory for inquirers to discover. To be wholly egalitarian about all carvings of the world would be to give away far too much to those who view inquiry as to the investigation of our own minds. (ibid., p. 399)

Capturing the logical structure of reality is more than a matter of discovering the right interpretation of the quantifiers. There are other logical expressions, such as sentence connectives and predicate modifiers. And, crucially, for Sider there are also predicates that carve nature at the joints, by virtue of referring to genuine “natural” properties. It is not for nothing that Sider's more recent book-length defense of his ontological realism is titled *Writing the Book of the World*, for here he defends this claim: “The world has a distinguished structure, a privileged description . . . There is an objectively correct way to ‘write

the book of the world.' " It goes without saying that this super-strength realism will not be to everyone's taste. But the Leibnizian dream of a perfect language that is in total harmony with reality has found adherents in analytic philosophy since its earliest days; Wittgenstein's *Tractatus* is perhaps the most famous attempt to characterize such a language—and the reality to which it corresponds.¹⁹ Irrespective of its ultimate fate, Sider's program is very much in the same vein.

Half a century ago, in his contribution to the Schilpp volume devoted to Carnap's work, Nelson Goodman wrote of the "evil days" currently being enjoyed by the *Aufbau*:

The *Aufbau* is a crystallization of much that is widely regarded as worst in 20th century philosophy. It is an anathema to anti-empirical metaphysicians and to alogical empiricists, to analytic Oxonians and to anti-analytic Bergsonians, to those who would exalt philosophy above the sciences and to those who would abolish philosophy in favor of the sciences. A good part of current polemical writing in philosophical journals is directed against views found in virulent form in the *Aufbau*. The *Aufbau* stands preeminent as a horrible example. (1963, p. 545)

Goodman was writing in the early 1960s, but in the years that followed this attitude to Carnap's first major work continued to be widely held. But for some at least—and not least Goodman himself—the appeal of the constructive program remained very much alive. And the appeal lives on. David Chalmers' *Constructing the World*—based on his 2010 John Locke Lectures in Oxford—is devoted to vindicating a version of Carnap's project:

In many ways, Carnap is the hero of this book. Like the other twentieth-century logical empiricists, he is often dismissed as a proponent of a failed research program. But I am inclined to think that Carnap was fundamentally right more often than he was fundamentally wrong. I do not think that he was right about everything, but I think that many of his ideas have been underappreciated. So one might see this project, in part, as aiming for a sort of vindication.

The title of this book is a homage to Carnap's 1928 book *Der Logische Aufbau der Welt* . . . one can see the current book as trying to carry out a version of Carnap's project in the *Aufbau*: roughly, constructing a blueprint of the world, or at least providing a vocabulary in which such a blueprint can be given. The aim is to specify the structure of the world in the form of certain basic truths from which all truths can be derived. To do this, I think one has to expand Carnap's class of basic truths and change the derivation relation . . . But with these changes made, I think

that the project is viable and that some of the spirit of the *Aufbau* remains intact. (2012, pp. xvii–xviii)

In the *Aufbau* Carnap held that all expressions are definable in terms of a more limited range of expressions, but since finite definitions look to be unavailable for many ordinary language statements, we need an alternative. At the center of Chalmers' constructive program is the "*a priori* scrutability" thesis. According to the latter, the complete set of truths about our world could be deduced from a class of basic truths, involving only a restricted range of concepts. Give a Laplacean demon the basic truths, and it could work out all the rest, using reason alone. What are these basic truths? One plausible scrutability base consists of all the microphysical truths, all the phenomenal truths, and all the indexical truths. But there are other options, some of which Chalmers explores in considerable detail. As for which is *the* best, he remains noncommittal: "It is clear that there are many possible *Aufbaus*. They vary with their basic class of expressions and their mode of construction. Some are closer to Carnap than others. Some are more successful than others" (2012, p. 246). What is not in doubt, for Chalmers at least, is that *some* version of something along the same lines as the *Aufbau* could be successful, and philosophically illuminating in many ways.²⁰

In a very different vein, Derek Parfit recently published *On What Matters* (2011), a massive two-volume (and 1,365-page) book on ethics. The work is noteworthy for reasons other than its sheer size: it is a hugely ambitious attempt to put an end to centuries of ethical disputes, and put the entire field on new foundations. As is well known, there are a great many ethical theories, but they fall into more general categories: Kantianism, consequentialism, and contractualism (the tradition of Hobbes and Locke, and more recently Rawls and Scanlon). These approaches are normally viewed as rivals—not surprisingly, given that Kantians hold that from a moral perspective it is the motive for an action that matters, whereas consequentialists hold that it is the effects of the action which are crucial. After a painstaking analysis, Parfit first identifies (what he sees as) the *best* or most defensible versions of Kantianism, contractualism, and consequentialism. He then argues that these theories in fact converge: they agree on which actions are right and which actions are wrong. Parfit is thus led to his "triple theory" of morality, a tripartite supreme principle that combines an element of each of the three supposedly competing theories: "an act is wrong just when such acts are disallowed by the principles that are optimific, uniquely universally willable, and not reasonably rejectable." An early draft of the book was called *Climbing the Mountain*; here is why:

It has been widely believed that there are such deep disagreements between Kantians, Contractualists and Consequentialists. That, I have

argued, is not true. These people are climbing the mountain on different sides. (2011, vol. 1, p. 419)

The Triple Theory, if correct, would itself constitute an enormous advance in moral theory, but Parfit is by no means done. Like Moore before him, Parfit rejects reductive “naturalistic” metaethical theories. He too believes that moral judgments can be objectively true or false, in much the same way as mathematical statements can be true or false. We can see that $2 + 2 = 4$, we can see that we each have a reason to avoid suffering severe pain in the future; there are self-evident normative truths concerning reasons, just as there are self-evident truths concerning numbers. In the final part of the book he engages in a lengthy defense of this objectivist position, and the stakes here are high: if the case for objectivism fails, Parfit maintains, *nothing* at all can matter, nihilism is unavoidable. Here Parfit is not just exhibiting continuity with an earlier analytic philosopher, he is extending a tradition in moral philosophy that stretches back as far as Plato. One of the strongest objections against the objectivity of ethics is the (supposed) fact of moral disagreement: there are far more moral disputes than there are disputes about elementary arithmetic. Hence the importance of the first part of Parfit’s argument. If he is right, there are far fewer serious moral disagreements than has previously been thought.

Taking a step or two back, the projects we have been looking at latterly are very different, but they also share some distinctive features. They are all highly ambitious, and they all engage with fundamental issues. Moreover, they are all seeking to defend or elaborate doctrines to that at least some of the founders of the analytic movement would have subscribed. Of course, quite what the verdict of posterity on them will be we cannot, as of now, know. But one thing at least is clear. The notion that at the start of its second century analytic philosophy is suffering from, or paralyzed by, a serious crisis of confidence looks to be well wide of the mark.

31 Coda A: What is Analytic Philosophy?

*Barry Dainton and
Howard Robinson*

We have told something of the story of “the analytic movement” as it ran from Russell, Moore, and Frege to the present day, but inevitably there are questions that remain unanswered. The following thought in particular might puzzle a reader: given the diversity to be found within it—some of which we have seen in the preceding pages—what, if anything, unites this tradition? Does analytic philosophy still exist as a single movement in any meaningful sense? There is no easy or straightforward answer to this question, but we believe there are nonetheless grounds for answering in the affirmative.

It is certainly true that anyone seeking a feature that is common to all of analytic philosophy—or its practitioners—is likely to be disappointed. The history reveals that several of the originators of the tradition were concerned with the foundation of mathematics. But in contemporary analytic philosophy this is an often technical and somewhat esoteric subdiscipline, and many of the issues that today’s philosophers are most concerned with are largely independent of the philosophy of mathematics. Is an unwavering commitment to clear and transparent argument a distinctive and universal characteristic of analytic philosophers? Although many do strive for maximum clarity, this is by no means always the case: Wittgenstein is the most striking counterexample, but there are others. One might think that a commitment to *analysis* is bound to be what is most distinctive about analytical philosophers. But as the historical overview demonstrates, there are very different conceptions of analysis to be found within the analytic tradition. For some “analysis” means an investigation into concepts, a search for nontrivial, necessary, and sufficient conditions for something to count as (say) a perceptual experience, or a free action. For those impressed by Russell’s theory of descriptions, analysis is a matter of revealing the true but concealed logical form underlying ordinary language statements. For others it is a matter of carefully studying the way expressions are actually used in ordinary language, with a view to dissolving rather than solving philosophical problems.

Of course, there are other possibilities. Is a commitment to a hard-headed empiricism a reliable mark of the analytic philosopher? How about nominalism? Or physicalism? Hardly: there are prominent defenders of (versions of) rationalism, a long tradition of belief in abstract objects (such as numbers). Also, dualism and idealism—not to mention panpsychism—have all been defended in recent years.¹ One contemporary young philosopher with impeccable analytic credentials now argues that the Hegelians had a point and that “the whole” is the primary substance and all forms of atomism are mistaken.

If we look to what leading analytical philosophers have themselves said in response to the question “What is analytic philosophy?” we find disagreement here too. Some philosophers hold that the correct philosophy has its roots in Frege and the resultant philosophy of logic and language. Michael Dummett, for example, claims that analytical philosophy was born when “the linguistic turn” was taken, and this turn was taken first by Frege:

Only with Frege was the proper object of philosophy finally established: namely, first, that the goal of philosophy is the analysis of the structure of *thought*; secondly that the study of *thought* is to be sharply distinguished from the psychological process of thinking; and, finally, that the only proper method for analysing thought consists in the analysis of *language*. (Dummett 1978, p. 458)

But the picture is more complicated than Dummett paints. In the nineteenth century Frege was by no means the only one following this path; so were Bolzano, Brentano, and Meinong. Indeed, Bolzano’s (1837) *Theory of Science* has a strong claim to be the first work in the analytic tradition, and the latter’s posthumous *Paradoxes of the Infinite* (1851) would have an important influence on mathematicians such as Dedekind and Cantor.²

Timothy Williamson expresses a general discontent with the narrowness of Dummett’s conception when he says that, from the perspective of many contemporary philosophers

... the conceptual turn and *a fortiori* the linguistic turn look like wrong turnings. It is pointless to deny that such philosophers are “analytic,” [sic] for that term is customarily applied to a broad, loose tradition held together by an intricate network of causal ties of influence and communication, not by shared essential properties of doctrine or method: what do Frege, Russell, Moore, Wittgenstein, Carnap, Ayer, Quine, Austin, Strawson, Davidson, Rawls, Williams, Anscombe, Geach, Armstrong, Smart, Fodor, Dummett, Wiggins, Marcus, Hintikka, Kaplan, Lewis, Kripke, Fine, van Inwagen and Stalnaker all have in common to distinguish them from all the non-analytic philosophers?

Many who regard the linguistic and conceptual turns as serious mistakes have ties of influence and communication that put them squarely within that tradition. "Analytic philosophy" is a phrase in a living language; the attempt to stipulate a sense for it that excludes many of the philosophers just listed will achieve nothing but brief terminological confusion. (2007, p. 21)

When it comes to methodology, in his homiletic "Must Do Better" Williamson berates contemporary analytic philosophers for their failure to maintain the highest of standards, and stresses the role of formal semantics in solving or clarifying major philosophical problems: "The attempt to provide a semantic theory that coheres with a given metaphysical claim can therefore constitute a searching test of the latter claim, even though semantics and metaphysics have different objects" (2007, pp. 284–5). On the basis of this passage alone it would not be unfair to characterize Williamson as a *methodological semanticist*, even if he is not a *ideological semanticist*. But he immediately goes on to acknowledge that a more liberal methodological stance is required:

Discipline from semantics is only one kind of philosophical discipline. It is insufficient by itself for the conduct of philosophical inquiry, and may sometimes fail to be useful, when the semantic forms of the relevant linguistic constructions are simple and obvious. But when philosophy is not disciplined by semantics, it must be disciplined by something else: syntax, logic, common sense, imaginary examples, the findings of other disciplines (mathematics, physics, biology, psychology, history, etc.) or the aesthetic evaluation of theories (elegance, simplicity, etc.). (ibid.)

It seems as though there is no simple or straightforward advice to be given to analytic philosophers who are eager to do better. We saw earlier that Williamson has been very impressed by the achievements of recent philosophy—"In many areas of philosophy, we know much more in 2007 than we knew in 1957; much more was known in 1957 than in 1907; much more was known in 1907 than in 1857" (2007, p. 280)—but the examples he goes on to cite all relate to the more technical areas of the subject, such as modal logic, semantic theories of truth, and the like. Valuable though these advances are, one may be skeptical about how far they take us. Modal logic cannot answer the metaphysical question of the status of possible worlds by telling us whether they are concrete and real, as Lewis thought, or quasi-real, in a Blackburnian way, or simply fictional, as Rosen and Armstrong have suggested, or Plantinga's "books" of propositions.³ Nor are all sophisticated philosophers convinced that the Tarskian schema tells us a great deal about truth on its own.

The catholicity of the tradition that Williamson illustrates with his list of names fits with the common experience of the profession. The editors of this volume studied in Oxford, one in the 1980s and the other in the 1960s. Oxford produced more professional philosophers in the “Anglo-American tradition” than any other university, but even in preceding decades, when a version of ordinary language philosophy still reigned, the education offered to students was never focused exclusively, nor even primarily, on the period we discuss in the history above. The central paper in the largest philosophy school (Politics, Philosophy and Economics) was based on the major texts of Descartes, Locke, Berkeley, and Hume, and was not taught as history of philosophy, but as a major route into many modern philosophical questions. The second largest school, “Classical Greats,” centered on Plato’s *Republic* and Aristotle’s *Nicomachean Ethics*, and even there the emphasis was by no means exclusively historical. What one might designate as the more narrowly analytic tradition tended to fall under a paper entitled “philosophical logic”—largely concerned with transformational analysis, and issues in the philosophy of logic and language—and views as to how central this is to major philosophical issues as the mind-body problem, free-will, our knowledge of the external world, and the nature of right and wrong varied greatly, as they still do.

So drawing a clear boundary between analytic and nonanalytic philosophy and philosophers is not easy. But distinguishing features do exist. As a useful first step we can distinguish what one might call *analytical philosophy in the narrow sense* and *analytical philosophy in the broad sense*. That there is a distinction along these lines is something most analytic philosophers would agree upon—indeed, it obtains through the whole period we have been discussing—even if they would also concur that it is difficult to pin down and characterize in an entirely satisfactory manner.

Philosophers conducting analytic philosophy in the narrow sense devote the bulk of their efforts to core areas of the subject, such as the philosophy of language, the philosophy of logic or mathematics, and the philosophy of science or mind. Since someone who concentrates on just one of these areas is more tightly focused than someone who works in two or more, “narrowness” in this sense is obviously a matter of degree. However, since Hegel and Heidegger, and their contemporary followers also have their own philosophies of logic, language, mind, and science, this criterion alone will not take us very far.

The situation is improved by a methodological supplement: it is *also* characteristic of the narrow analytic philosopher to privilege one of the modes of analysis mentioned above. Hence such a philosopher may well believe (say) that progress on a range of important issues in the core areas of the subject is most likely to be achieved by revealing the underlying logical form of the relevant classes of statement—or alternatively, by paying attention to easily overlooked

facts about ordinary usage, or employing decompositional conceptual analysis. Since it is possible to privilege more than one mode of analysis—for example, one might believe some problems will be resolved by transformational analysis, others by ordinary language dissolution—this criterion can also be satisfied to different degrees. A philosopher who subscribes to just one mode of analysis is narrower than one who subscribes to two or three.

It is obvious that Wittgenstein, both early and late, comes out as narrow by this multidimensional criterion, as do Russell, Carnap, Austin, Quine, and Davidson. Moreover, those working primarily on issues that lie outside the “core” areas—for example, in ethics, political theory, or aesthetics—but who employ one or more of these methods of analysis will also now count as analytic philosophers in the narrow sense. This is clearly the right result. Political philosophers who employ logic and game theory, or aestheticians who draw on modal logic, are evidently analytic philosophers in the narrow sense if anyone is.

In contrast, the broad analytic philosophers do not confine their attention to issues in the “core” areas of logic, language, mind, and science, and they do not rely heavily on any particular methodological approach. The work of C. D. Broad, for example, is not analytic in the narrow sense, nor that of H. H. Price, but Broad’s *The Mind and its Place in Nature* has returned to fashion as a “great book” in the analytic tradition, and Price’s *Perception* has never ceased to be a major reference point, even for those who disagree with it bitterly. Many other examples could be given, for example, Sellars’ *Science, Perception and Reality*, Dennett’s *Consciousness Explained*, or Parfit’s *Reasons and Persons*.

As for whether it is possible to provide a usefully informative characterization of what makes for an analytical philosopher in the broad sense, we might say something along these lines: what unifies them is the belief that most of the concepts of interest to philosophers are, in one way or another, complex both in themselves and in their relations to other concepts, and it is necessary to spend a considerable amount of time and thought coming to grips with this complexity before blithely using them in service of some favored project. This process of coming to grips with the complexity may very well show that the project one had in mind was illegitimate. But such a very generic notion gives analytic philosophy a very wide ambit: Aristotle, Leibniz, Hume, and Kant probably qualify.

The only obvious way of restricting the scope of the “broadly analytic” so that it only applies to philosophers who would *call* themselves analytic is by adding a historical condition. Dean Zimmerman has recently suggested that

the distinctive thing about analytic philosophers is that they see themselves as the rightful heirs of Russell and Moore, or of philosophers who saw themselves as the rightful heirs of Russell and Moore, or . . .

“Analytic,” so understood, is an adjective grounded, rather loosely, in the way philosophers think about their debts to their predecessors active at the beginning of the twentieth century. To be an analytic philosopher is to accept a version of the history of philosophy according to which the heroes at the beginning of the last century were Frege, Russell, and Moore—not Bradley, Bosanquet, and Bergson. It is to admire the philosophical impact of the analytic revolutionaries, and to hope to be a similar “force for good” in one’s own time. (2004, p. xv)

Appealing to historical relationships successfully separates Aristotle and Locke from Carnap and Kripke, but we are not yet done. As Zimmerman also points out, in trying to define analytic philosophy “in its broadest sense” we can appeal to both historical connections and *self-definition*. In their own thinking about themselves and their tradition analytic philosophers draw heavily on a contrast with so-called continental philosophy. Analytic philosophers in the broad sense will certainly view themselves as following in the footsteps of Russell, Moore, Frege, and Wittgenstein, but they will also view themselves as most definitely *not* doing the sort of philosophy that has dominated much of continental Europe since the 1920s (see below for more on the distinction between the analytic and continental traditions).

There are some philosophers who think that the fragmentation or dissipation of the analytic tradition in the narrow sense means that the distinction between analytic and continental philosophy is now outdated. For the reasons outlined in the previous section we believe that it is a mistake to think that analytic philosophy, even in the narrow sense, has exhausted itself.

32 Coda B: Analytic versus Continental

Howard Robinson

Toward the end of his illuminating and entertaining *What is Analytic Philosophy?* Hans-Johan Glock considers the suggestion that there is no longer any reason to distinguish analytic and continental philosophy and that they should try to come together. He argues that this is not the case:

Philosophy is not about sharing doctrines, but about a rational and civilised debate even about one's own cherished assumptions. Such a debate remains easier among analytic philosophers than between analytic and continental philosophers. . . .

Admittedly, there are now very competent expositors of continental philosophy, mostly Anglophone philosophers with some analytic background.¹ But the genuinely continental and original voices in that field, in so far as any remain, strike me to be as obscure as ever . . .

While there may be a premium on reconstituting philosophy as a unified sphere of discourse, this must not go at the expense of rigour, clarity, scholarship and intellectual honesty. (pp. 259–60)

And with this I fully concur. The “contrastive” core of the broad analytic conviction could be expressed by saying that something went seriously wrong with the practice of philosophy with, and under the influence of, the work of Hegel. What went wrong was not a matter of doctrine, as such—at least, not if that means idealism or the priority of the “whole”—but of the style of philosophical discourse itself. The problem analytics have with continentals is not a matter of *disagreement*—they have those with each other with relish—but of failure to engage. Nor do analytics in general want to accept that this is just a matter of “difference of tradition”—people doing things in different, equally valid ways.

The analytic complaint against continental philosophers is not merely that it is hard to communicate with them argumentatively—in the benign sense of that term, as it applies, almost uniquely, within philosophy—but the suspicion is that they do not communicate so much in that way with each other. This is because there is not such a set of free-standing problems, as expressed in the chapters of this volume, which can be tackled in their own right and not only in the context of X's or Y's system.² It is the degree of the autonomy of the issues that, to a large extent, marks out the difference from continental philosophy, and, in this, modern continental philosophy—despite boasting of its connection with the history of philosophy—differs from the bulk of philosophy before Kant and Hegel.

I put this point about the relative solipsism of Continental practice to a colleague who specializes in Continental philosophy, expecting a rebuttal. To my surprise he conceded that it was—distressingly—true. On planning a graduate course on Continental aesthetics he chose a reader with 19 chapters by different authors. On teaching the course he found that, though each chapter was interesting, there was no overlap or common themes, and no developing discussion was possible.

This view is supported—at least with regard to recent French philosophy—by Gary Gutting, who is equally at home in either tradition. He says of writers such as Deleuze, Derrida, and Levinas:

My concern, however, is about the obscurity that arises because authors do not make a sufficient effort to connect their novel concepts to more familiar (even if technical) concepts that would allow an informed and conscientious reader to make an assessment of their claims. The result is writing that is *hermetic* in the sense that it effectively cuts itself off from the very issues of common concern that it is trying to address. . . .

This is why, in contrast to the world of analytic philosophy, there has been so little profitable, progressive debate about the views of the major French philosophers. In a country with so many teachers and students of philosophy, there should be a strong tradition of lively discussion in articles and colloquia. Instead, the primary medium of expression has been a series of forbidding *magna opera*, so difficult as to be almost untouchable by detailed criticism. (2011, p. 200)

How this rhetorical solipsism of the great masters of modern continental thought came about and operates is a further issue, and not an easy one. I would suggest two sources.

The first is broadly a matter of intellectual content. Although it was suggested above that the “continental” deviation from the mainstream began with Hegel, its roots are probably Kantian. Prior to Kant, philosophers worked in

a common workspace, which included an objective notion of rationality and no discontinuity (which does not mean no difference) between philosophy and natural science. Kant privatized rationality by making it the creature of the structure of the human mind and without application beyond the range of our experience. Of course, the Kantian categories were meant to be common to all humanity, but once the debate came to be about different views about the *a priori* structure of the human mind, proponents of each of the subsequent neo-Kantian systems were enclosed within its own world of necessary *a priori* psychology. Natural science became an outcast, and no longer a common source of reflection and information; common beliefs and our ordinary concepts lost their universal currency as objects for analysis, and there was no common grounding in logic. Philosophers ceased to be people with different views engaged in a common activity of debate on shared vital issues, but expositors and advocates of more or less enclosed systems.

Gutting has an explanation that is consistent with this. Kant believed that our concepts led to contradiction when applied beyond the empirical. Hegel brought these contradictions home, saying that they applied throughout, but by the dialectic they would finally be resolved. Much of modern French philosophy, according to Gutting, rests on the idea that contradictions can never be avoided—in Derrida's terms, every concept deconstructs itself. This makes straightforward rational discussion impossible, and one can only play with the conceits that lie in the interstices of concepts.

The second point is that this situation was no doubt greatly aggravated by the sociology of continental universities. Leading figures did not engage in regular close debate and discussion with each other and with students at close quarters, but each master tended to pontificate to large numbers of adoring disciples. An unhealthy self-indulgence was almost unavoidable. This can be contrasted with the sociology of the subject in its Anglo-Saxon countries. A. J. Ayer, in Oxford, for example, ran two discussion groups, one for senior faculty, one mainly for undergraduates and graduates—the latter, his famous “informal instruction.” The senior group, by invitation, met about twice a term where people read each other papers over drinks. The members included Peter Strawson, Michael Dummett, John Mackie, David Wiggins, Simon Blackburn, Gareth Evans, John McDowell, Derek Parfit, Christopher Peacocke, John Foster, Michael Woods, as well as distinguished visitors, such as Quine, Davidson, and Goodman, when passing through. It is said that only once in 30 years was there a personal dispute that led to someone (not one of those above) walking out. A more diverse group of major philosophers can hardly be constructed and such a phenomenon within the continental tradition can hardly be imagined. Similar, if less high-powered, groups are normal within the subject. Once again, it is the availability of the Socratic marketplace that marks and makes the difference between the analytic and continental traditions.

If one wants to grasp the contrast between the analytic and the continental spirits, one should perhaps look at what those analytic philosophers located on the continent say about the difference. They were until recently a besieged minority and they are not disposed to compromise. The *European Society for Analytic Philosophy* characterizes itself on its Website as follows:

Analytic philosophy is characterized above all by the goal of clarity, the insistence on explicit explanation in philosophy, and the demand that any view expressed be exposed to the rigours of critical evaluation and discussion by peers. (Filosofia.dafist.unige.it/esap/)

One might think that these are features of any good philosophy. Indeed, Angsar Beckermann, a former president of the *Gesellschaft für Analytische Philosophie*, after characterizing modern analytic philosophy in more or less the same way as European Society of Analytic Philosophy (ESAP), concludes: "And in my view these are indeed the distinguishing features of good philosophy" (2004, p. 12: translated and quoted by Glock, 2006).

It would be disingenuous not to say clearly that analytic philosophers, on the whole, think continental philosophy is plain bad philosophy and lacking the basic standards of philosophical argument and even intellectual integrity, at least in some of its most recent forms.

Notes

Part III: New Directions in Analytic Philosophy

- 1 <http://philpapers.org>. At the time of writing (May 2012) there are some 11,000 entries for Epistemology, 12,000 for Metaphysics, 7,000 for Philosophy of Action, 15,000 for Philosophy of Language, 23,000 for Philosophy of Mind, 11,000 for Philosophy of Religion, 4,000 for Meta-Ethics, 8,000 for Normative Ethics, 13,000 for Logic and Philosophy of Logic, 3,000 for Philosophy of Mathematics, 10,000 for Philosophy of Science (General and Physical Sciences), and 18,000 for Philosophy of Cognitive Science—figures that give an approximate indication of the areas in which most work is being done.
- 2 Some of the influential earlier publications include: Earman and Norton 1987; Butterfield 1988; Norton 1988; Maudlin 1990; Earman 1992.
- 3 See Earman 1995, 2002; Maudlin 2004; and Rickles et al. 2006.
- 4 As always there are exceptions, see (for example) Hacker and Bennett 2003 for a Wittgensteinian critique of methodological naturalism in the philosophy of mind.
- 5 See Dainton 2000/2006, Tye 2003, and Bayne 2010. For a survey of recent work on the temporal character of experience, see <http://plato.stanford.edu/entries/consciousness-temporal/>
- 6 Influential representationalist writings include Harman 1990 and Tye 1995.
- 7 For a useful introduction to the debate see Kind 2010; also: <http://plato.stanford.edu/entries/consciousness-representational/>
- 8 For an overview of these works see Dainton 2012.
- 9 The expression “cognitive phenomenology” is also used in this connection, and more or less equivalently.
- 10 In *Wittgenstein on Rules and Private Language* (1982) Kripke argues the totality of facts—pertaining to one’s entire mental and physical history—determine whether one means “plus” or “quus” by *plus*, where quus is a deviant function that delivers the answer five to any addition involving numbers greater than 50. For the sake of the argument we assume you have never added numbers greater than 50 before (for more on Kripke’s interpretation of Wittgenstein see Chapter 18). In his *Reason, Truth and History* (1981) Putnam argues that even if we know the truth of the sentence “The cat is on the mat” in every possible world we will not know what the terms “cat” and “mat” refer to—and the same applies for all other referring expressions.
- 11 For example, Sider (forthcoming) “Against Parthood,” tedsider.org/papers/nihilism.pdf; Dorr and Rosen 2002 and Cameron 2007.
- 12 See, for example, Richardson 1998, Friedman 1999, and also Richardson and Uebel 2007. According to the “new” interpretation of the *Aufbau*, Carnap’s primary concern there was not with carrying out a successful phenomenistic reduction, but

- rather establishing that *objectivity* in science—or formulations of scientific theories—was to be secured by eliminating everything that is not structural, or as he put it (§16) “science wants to speak about what is objective, and whatever does not belong to the structure but to the material . . . is, in the final analysis, subjective.” A thesis that is very much alive in contemporary philosophy of science. For an overview of interpretations of the *Aufbau* see Pincock 2009. For an overview of contemporary “structural realism” see, <http://plato.stanford.edu/entries/structural-realism/>
- 13 For a survey of this work see Knobe and Nichols 2008; also Paul 2010.
 - 14 It is not for nothing that McGinn’s book is subtitled “Games, Names and Philosophy”: in his *Preface* he relates how he had come across a book—*The Grasshopper: Games, Life and Utopia* by Bernard Suits—where “the author, speaking in the persona of his wise insect, announces his intention of defining the concept *game*. I shook my head: *that* isn’t going to fly (or hop). For a good thirty-five years I had followed the conventional wisdom that Wittgenstein had shown that the concept *game* is not definable by means of necessary and sufficient conditions. Moreover, when I first read the purported definition I felt unconvinced, suspecting counterexamples by the bushel. The book pressed on, bringing up and defusing one putative counterexample after another . . . After a second reading, and then a third I gave up the fight: games had been defined! . . . Now the search for definition—for conceptual analysis—seemed like not such a misguided enterprise, even for the most recalcitrant of cases” (2012, p. vii).
 - 15 This was the first of three important works on Frege, the others being *The Interpretation of Frege’s Philosophy* (1981) and *Frege Philosophy of Mathematics* (1991). For an introduction to Dummett’s own distinctive philosophical position see the introduction to his *Truth and Other Enigmas* (1978).
 - 16 *Wittgenstein: Understanding and Meaning Volume 1* (1980, revised 2nd edn 2005); *Wittgenstein: Rules, Grammar and Necessity, Volume 2* (1985). The final two volumes were completed by Hacker alone: *Wittgenstein: Meaning and Mind, Volume 3* (1990b) and *Wittgenstein: Mind and Will, Volume 4* (1996b).
 - 17 See, for example, Wright 1983; for an overview of several neo-logician approaches see Linsky and Zalta 2006.
 - 18 Or as Gaskin puts it: “. . . if the generation of Bradley’s regress is thus implicated in the metaphysical origin of language, linguistic idealism will tell us that it also sustains the very existence of the world in the philosophically hygienic sense of that word—the totality of (true and false) propositions at the level of reference” (2008, p. 420). See Gaskin 2010 for a *précis* of the book and critical responses to it. Schaffer’s “The Internal Relatedness of All Things” (2010) is another example of a contemporary philosopher returning to an issue at the forefront of debates in the early days of the analytic movement. Schaffer argues, *contra* Russell and Moore, that there may well be powerful arguments in favor of a doctrine primarily associated with idealists such as Bradley, that reality consists of a single whole, one whose parts are all essentially interdependent.
 - 19 Writing of his hopes for his own perfect language, Leibniz wrote the following to the Duke of Hanover in 1679: “. . . my invention uses reason in its entirety and is, in addition, a judge of controversies, an interpreter of notions, a balance of probabilities, a compass which will guide us over the ocean of experiences, an inventory of things, a table of thoughts, a microscope for scrutinizing present things, a non-chimerical cabal, a script which all will read in their own language; and even a language which one will be able to learn in a few weeks, and which will soon be accepted amidst the world. And which will lead the way to the true religion everywhere it goes.” The conviction (or hope) that such a language exists is by

- no means confined to analytic philosophy, and long pre-dates Leibniz, and—quite probably—Western philosophy. For more on this theme see Eco 1995.
- 20 Chalmers is not alone in viewing the *Aufbau* as a promising start: see also Hannes Leitgeb, “New Life for Carnap’s *Aufbau*?” (forthcoming).

Chapter 31

- 1 John Foster’s *The Case for Idealism* (1982) is perhaps the leading assault on the reality of the physical world by a respected analytic philosopher. Foster has also defended dualism—see *The Immaterial Mind* (1991).
- 2 See Coffa 1991 and Simons 1999. David Bell’s “The Revolution of Moore and Russell: A Very British Coup?” (1999) begins thus: “Why did analytic philosophy emerge first in Cambridge, in the hands of G. E. Moore and Bertrand Russell, and as a direct consequence of their revolutionary rejection of the philosophical tenets that form the basis of British Idealism? . . . the answer I shall try to defend is: it didn’t.” Bell goes to argue “Moore . . . is best seen as the major, though by no means the first, British participant in an existing debate whose other participants included Ward, Stout, Russell, Meinong, Stumpf, Husserl, Twardowski and Brentano. Many of the terms and goals of this debate originated in Germany, during the 1870s, in the attempts by philosophers, physiologists, theologians and others to come to terms with, and contribute to, the emergence of psychology as a discipline in its own right. Russell, too, during the period between 1899 and 1903 is best seen as engaging with issues and innovations associated, on the one hand, with the logico-mathematical works of Dedekind, Schröder, Cantor, Klein, Riemann, Helmholtz, Bolzano, Peano and Frege, and on the other hand, with the contributions to psychology and philosophy made by Brentano, Meinong, Ward, Stout, Fechner, Helmholtz and, of course, Moore” (p. 208). For more on the influence of Bolzano, Brentano, and Meinong—and a useful introduction to the influence of Austrian philosophy on twentieth-century philosophy as a whole—see Textor 2006.
- 3 See Lewis 1986, Blackburn 1987, Armstrong 1989, Rosen 1990, and Plantinga 2003.

Chapter 32

- 1 This remark of Glock’s is attested by the case of the new *Oxford Handbook of Continental Philosophy*, edited by Brian Leiter and Michael Rosen. Only one contributor is employed on the mainland of Europe—the rest work in the United States or Britain: and the one based on the Continent has strong Oxford connections.
- 2 Gutting also suggests that the work of recent French philosophers is superior to the bulk of recent analytic work by virtue (among other things) of their willingness to question the limitations of conceptual thought, and develop radically new concepts in their attempt to “think the impossible.” It should be pointed out that analytic philosophers have not ignored these issues. In recent decades there has been much debate in analytical circles as to whether “non-conceptual content” features in perception. And “the impossible” has not been ignored, far from it: see the *Stanford Encyclopedia of Philosophy* entry on “Impossible Worlds” <http://plato.stanford.edu/entries/impossible-worlds/>

Chronology

- 1837 Bolzano's *Wissenschaftslehre* (Theory of Science) is published — considered by some to be the first work in analytic philosophy.
- 1879 Frege publishes *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens* (translated as *Concept Script, a Formal Language of Pure Thought Modelled upon that of Arithmetic*).
- 1884 Frege: *Die Grundlagen der Arithmetik: eine logisch-mathematische Untersuchung über den Begriff der Zahl* (translated as *The Foundations of Arithmetic: A logico-mathematical Enquiry into the Concept of Number*).
- 1892 Frege's "Über Sinn und Bedeutung" (translated as *On Sense and Reference*).
- 1893 Frege's *Grundgesetze der Arithmetik* (*The Basic Laws of Arithmetic*).
- 1894 Frege reviews Husserl's *Philosophy of Arithmetic* — the review is largely hostile: "In reading, this work I was able to gauge the devastation caused by the influx of psychology into logic . . . The mistakes which I thought it my duty to show reflect less upon the author than they are the result of a widespread philosophical disease."
- 1899 Moore's "The Nature of Judgement."
- 1901 Discovery of Russell's eponymous (set theoretic) paradox.
- 1903 Russell's *The Principles of Mathematics*; Moore's *Principia Ethica* and "The Refutation of Idealism."
- 1905 Russell's "On Denoting" is published in *Mind*.
- 1910–13 Russell and Whitehead publish *Principia Mathematica*.
- 1911 Russell and Bergson clash at a meeting of the Aristotelian Society — the first of several (Russell later remarks that "intuition is at its best in bats, bees and Bergson" — thus marking his rejection of Bergson's anti-intellectualism).
- 1912 Wittgenstein starts to study with Russell at Cambridge; Russell's *The Problems of Philosophy*.
- 1918–19 Russell lectures on Logical Atomism; Moore's "Internal and External Relations."

- 1921 Wittgenstein publishes *Tractatus Logico-Philosophicus*.
- 1922 Schlick given chair of Philosophy in Vienna, birth of the Vienna Circle; Moore's *Philosophical Studies*.
- 1921 Russell's *The Analysis of Mind*.
- 1923 Broad's *Scientific Thought*.
- 1925 Death of Frege; Broad's *The Mind and its Place in Nature*; Moore's "A Defense of Common Sense."
- 1928 Carnap's *The Logical Structure of the World and Pseudoproblems in Philosophy*.
- 1929 Wittgenstein returns to Cambridge, and resumes philosophy. An encounter at Davos takes place between Carnap, Cassirer, and Heidegger, and the division between (what would later be called) analytic and continental philosophy starts to open.
- 1931 Gödel's "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I" (On Formally Undecidable Propositions of *Principia Mathematica* and Related Systems); Ramsey's *The Foundations of Mathematics and Other Logical Essays*; Ryle's "Systematically Misleading Expressions."
- 1931–40 Dispersal of Vienna Circle and associates to the United States: Feigl (1931, from Vienna), Carnap (1935, from Prague), Hempel (1937, from Germany), Reichenbach (1938, from Germany), Bergman (1938, from Vienna), Tarski (1939, from Poland), and Gödel (1940, from Vienna).
- 1934 Popper's *Logik der Forschung* (published in English in 1959 as *The Logic of Scientific Discovery*).
- 1936 Murder of Schlick in Vienna; Ayer's *Language, Truth and Logic*.
- 1939 Moore's "Proof of an External World."
- 1944 Tarski's "The Semantic Conception of Truth and the Foundations of Semantics" published.
- 1945 Popper's *The Open Society and its Enemies*.
- 1947 Carnap's *Meaning and Necessity: A Study in Semantics and Modal Logic*.
- 1948 Russell's *Human Knowledge: Its Scope and Limits*; Quine's "On What There Is."
- 1949 Ryle publishes *The Concept of Mind*; Feigl and Sellars' *Readings in Philosophical Analysis*—a key teaching text for several decades.
- 1951 Wittgenstein dies; Quine's "Two Dogmas of Empiricism"; Goodman's *The Structure of Appearance* appears in print.
- 1951 Goodman's *Fact, Fiction and Forecast*.
- 1952 Hare's *The Language of Morals*.
- 1953 Posthumous publishing of Wittgenstein's *Philosophical Investigations*; Quine's *From a Logical Point of View*.

- 1956 Sellars' "Empiricism and the Philosophy of Mind."
- 1956–7 Austin's "Plea for Excuses," "Ifs and Cans."
- 1957 Chomsky's *Syntactic Structures*; Anscombe's "Intention"; Grice's "Meaning."
- 1958 The "Royaumont" conference takes place, where Ayer, Ryle, Quine, Austin, Hare, and Strawson meet Wahl, Breda, and Merleau-Ponty in an attempt to bridge the growing gulf between the analytic and continental movements (as they were now being called)—the attempt failed. Anscombe's "Modern Moral Philosophy" was also published.
- 1959 Strawson's *Individuals*; Kripke's first publication on modal logic (followed up in 1963, 1965); Malcom's *Dreaming*.
- 1960 Quine's *Word and Object*.
- 1961 Hart's *The Concept of Law*; Popper and Adorno cross swords at a meeting of the German Sociological Association.
- 1962 Austin's *How To Do Things With Words*; Kuhn publishes *The Structure of Scientific Revolutions*.
- 1963 Sellars' *Science, Perception and Reality*; Shoemaker's *Self-Knowledge and Self-Identity*; Popper's *Conjectures and Refutations*; Gettier's "Is Justified True Belief Knowledge?"
- 1965 Hempel's *Aspects of Scientific Explanation*.
- 1966 Strawson's *The Bounds of Sense*.
- 1967 Rorty's *The Linguistic Turn*; Davidson's "Truth and Meaning."
- 1968 Armstrong's *A Materialist Theory of the Mind*; Goodman's *Languages of Art*; Grice's "Utterer's Meaning, Sentence Meaning and Word-Meaning."
- 1969 Searle's *Speech Acts*.
- 1970 Death of Russell; Davidson's "Mental Events"; Sen's *Collective Choice and Social Welfare*.
- 1971 Rawls' *A Theory of Justice*; Thomson's "A Defense of Abortion."
- 1971–2 Kripke's "Identity and Necessity" and "Naming and Necessity" appear in article form; book *Naming and Necessity* appears in 1980.
- 1973 Dummett's *Frege: Philosophy of Language*; Williams' *Problems of the Self*.
- 1974 Nozick's *Anarchy, State and Utopia*; Plantinga's *The Nature of Necessity*; Sklar's *Space, Time and Spacetime*.
- 1975 Putnam's *Philosophical Papers* I and II, and "The Meaning of 'Meaning'"; Fodor's *The Language of Thought*; Grice's "Logic and Conversation"; Singer's *Animal Liberation*.
- 1976 Lewis' *Counterfactuals*.
- 1978 Dummett's *Truth and Other Enigmas*; Foot's *Virtues and Vices*.
- 1979 Nagel's *Mortal Questions*; Burge's "Individualism and the Mental."

- 1980 Richard Rorty publishes *Philosophy and the Mirror of Nature*; Davidson's *Essays on Action and Events*; Wiggins' *Sameness and Substance*; van Fraassen's *The Scientific Image*.
- 1981 Putnam's *Reason, Truth and History*; Anscombe's *Collected Papers I, II, and III*.
- 1982 Evans' *The Varieties of Reference*; Kripke publishes *Wittgenstein on Rules and Private Language*, offering a controversial new interpretation of the private language argument; Derrida engages with Austin and speech act theory in "Signature Event Context," Searle responds.
- 1984 Parfit's *Reasons and Persons* published; Davidson's *Inquiries into Truth and Interpretation*; Shoemaker's *Identity, Cause and Mind: Philosophical Essays*.
- 1985 Williams' *Ethics and the Limits of Philosophy*.
- 1986 David Lewis' *On the Plurality of Worlds* is published.
- 1990 Walton's *Mimesis as Make-Believe: On the Foundations of the Representational Arts*; Gibbard's *Wise Choices, Apt Feelings*; van Inwagen's *Material Beings*.
- 1991 Dennett's *Consciousness Explained*.
- 1992 Searle's *The Rediscovery of the Mind*.
- 1993 Rawls' *Political Liberalism*. Derrida awarded an honorary doctorate by the University of Cambridge—several analytic philosophers sign a letter of protest, where they state "Academic status based on what seems to us to be little more than semi-intelligible attacks upon the values of reason, truth and scholarship is not, we submit, sufficient grounds for the awarding of an honorary degree in a distinguished university."
- 1994 McDowell's *Mind and World*; Williamson's *Vagueness*; Brandom's *Making it Explicit*. The first *Towards a Science of Consciousness* conference held in Tucson.
- 1996 Chalmers' *The Conscious Mind: In Search of a Fundamental Theory*; the "Sokal hoax" perpetrated.
- 1998 Scanlon's *What We Owe to Each Other*.
- 2001 Death of David Lewis; publication of Sider's *Four-Dimensionalism*.
- 2010 David Chalmers' "Constructing the World" John Locke Lectures, Oxford.
- 2011 Parfit publishes *On What Matters*; death of Dummett.

Timeline of Individual Philosophers

- Bernard Bolzano (1781–1848)
- T. H. Green (1836–82)
- Ernst Mach (1838–1916)
- Charles Sanders Peirce (1839–1914)
- William James (1842–1910)
- F. H. Bradley (1846–1924)
- Bernard Bosanquet (1848–1923)
- Gottlob Frege (1848–1925)
- Alexius Meinong (1853–1920)
- Josiah Royce (1855–1916)
- Giuseppe Peano (1858–1932)
- John Dewey (1859–1952)
- Alfred North Whitehead (1861–1947)
- J. M. E. McTaggart (1866–1925)
- Bertrand Russell (1872–1970)
- G. E. Moore (1873–1958)
- W. D. Ross (1877–1971)
- Moritz Schlick (1882–1936)
- Otto Neurath (1882–1945)
- C. I. Lewis (1883–1964)
- C. D. Broad (1887–1971)
- Ludwig Wittgenstein (1889–1951)
- Rudolf Carnap (1891–1970)
- Friedrich Waismann (1896–1959)
- H. H. Price (1899–1984)
- Gilbert Ryle (1900–76)
- Alfred Tarski (1901–83)
- Karl Popper (1902–94)
- Alonzo Church (1903–95)

- Frank P. Ramsey (1903–30)
- Carl Hempel (1905–97)
- Kurt Gödel (1906–78)
- Nelson Goodman (1906–98)
- H. L. A. Hart (1907–92)
- Willard van Orman Quine (1908–2000)
- Charles Stevenson (1908–79)
- A. J. Ayer (1910–89)
- J. L. Austin (1911–60)
- Norman Malcolm (1911–90)
- Wilfrid Sellars (1912–89)
- Alan Turing (1912–54)
- H. P. Grice (1913–88)
- Arthur Prior (1914–16)
- Roderick Chisholm (1916–99)
- G. H. von Wright (1916–2003)
- Donald Davidson (1917–2003)
- J. L. Mackie (1917–81)
- G. E. M. Anscombe (1919–2001)
- R. M. Hare (1919–2002)
- P. F. Strawson (1919–2006)
- Philippa Foot (1920–2010)
- John Rawls (1921–2002)
- Ruth Barcan Marcus (1921–2012)
- Thomas Kuhn (1922–66)
- Michael Dummett (1925–2011)
- David M. Armstrong (1926–2014)
- Hilary Putnam (born 1926)
- Noam Chomsky (born 1928)
- Jaakko Hintikka (born 1929)
- Bernard Williams (1929–2003)
- Richard Rorty (1931–2007)
- Sydney Shoemaker (born 1931)
- Alvin Plantinga (born 1932)
- John Searle (born 1932)
- David Kaplan (born 1933)
- Amartya Sen (born 1933)
- Jerry Fodor (born 1935)
- Thomas Nagel (born 1937)
- Gilbert Harman (born 1938)
- Robert Nozick (1938–2002)
- Saul Kripke (born 1940)

Timeline of Individual Philosophers

- Thomas Scanlon (born 1940)
- Robert Stalnaker (born 1940)
- Bas van Fraassen (born 1941)
- David K. Lewis (1941–2001)
- Daniel Dennett (born 1942)
- Peter van Inwagen (born 1942)
- John McDowell (born 1942)
- Derek Parfit (born 1942)
- Crispin Wright (born 1942)
- Frank Jackson (born 1943)
- Simon Blackburn (born 1944)
- Tyler Burge (born 1946)
- Gareth Evans (1946–80)
- Hartry Field (born 1946)
- Kit Fine (born 1946)
- Robert Brandom (born 1950)
- Christopher Peacocke (born 1950)
- Timothy Williamson (born 1955)
- David Chalmers (born 1966)

A–Z of Key Terms and Concepts

1 ABSTRACT

So-called abstract objects (if they exist) are not material in nature, and not part of the spatio-temporal realm. Typical candidates: numbers, sets, universals, propositions. Existing things of the nonabstract variety are often called *concrete*—with no implication that they are composed of the building material of the same name.

2 ABSTRACTION

For empiricists such as Locke it is the process by which we acquire concepts by noticing features that are common to various instances. The term is also used in connection with procedure deployed by Frege when developing his account of number. A line *a* is parallel to a line *b* if and only if the direction of line *a* is identical with the direction of line *b*. In this manner, Frege argued, an “equality holding generally” can be equivalent to an *identity*. More generally, a Fregean *abstraction principle* is of the form $[a = b \leftrightarrow R(a,b)]$, where *R* is an equivalence relation.

3 ACTUALISM

The view that only the actual world is real. Although we may sometimes find it useful to talk in terms of other “possible worlds,” actualists hold that the latter are not really worlds, that is, they are not concrete entities in the manner of this, the actual world. See also *Modal Realism*, *Fictionalism*, and *Possible Worlds*.

4 ANALYSIS

Although a reliance on analysis is a distinguishing feature of analytic philosophy, different philosophers—or the same philosophers at different times—have meant different things by the term. For the early Moore and Russell (also Carnap), analysis was *decompositional*: we learn the real nature of a complex thing by revealing the more basic things of which it is composed. Or as Moore put it in “The Nature of Judgement”: “A thing becomes intelligible first when it is analysed into its constituent concepts.” In *Principia Ethica* Moore argues that the concept “good” is primitive and undefinable because it does not possess parts into which it can be decomposed. In a typical instance of *conceptual analysis*, an informative definition for a given concept *F* is supplied, one that—ideally—provides necessary and sufficient conditions for a state of affairs falling under *F*. Hence, as a first approximation, one might hold that “knowing that *P*” consists in the holding of the justified true belief that *P*. Since satisfactory analyses are supposed to hold across possible worlds, there is no need to confine oneself to the actual world in seeking counterexamples—hence the frequent appeal to exotic thought experiments in many areas of analytic philosophy.

Russell’s theory of descriptions exemplifies a different form of analysis: *logical* or *transformative*. A sentence such as “The present king of France is bald” appears to be expressing a simple subject-predicate proposition, but—according to Russell—its real form is very different, and hence heavily disguised by the surface grammar. More generally, proponents of this form of analysis hold that the true logical form of a sentence is revealed by translating it into the language of the predicate calculus (typically). This approach—pioneered by Frege and Russell, also embraced by the early Wittgenstein and Carnap—is viewed by many analytic philosophers as the paradigm of analysis. Many, but not all: ordinary language philosophers downplay the importance of transformative analysis, and place more emphasis on *connective analysis* as Strawson labeled it:

A concept may be complex, in the sense that its philosophical elucidation requires the establishing of its connections with other concepts, and yet at the same time irreducible, in the sense that it cannot be defined away, without circularity, in terms of other concepts to which it is necessarily related. (Strawson 1992, pp. 22–3)

5 ANALYTIC VERSUS SYNTHETIC

Analytic sentences are those that are true by virtue of the meaning of their constituent words, for example, “Bachelors are unmarried.” Sentences whose truth is not discoverable in this way—for example, “Bachelors are happier

than married men”—are synthetic. The distinction derives from Kant, who suggested (somewhat obscurely) that truths are analytic when the concept referred to by the predicate is “contained in” the concept of the subject, as is the case (he claimed) for “All bodies are extended.” Frege later refined this account, and argued that analytic truths are those that are true by virtue of logical laws and definitions. The doctrine that there is a real distinction between analytic and synthetic statements would later come under attack by Quine.

6 A PRIORI/A POSTERIORI

A statement (or proposition or sentence) is *a priori* if it is possible to discover whether it is true or false without consulting the world, either by direct observation or experiment. In order to verify or falsify an *a posteriori* statement it is necessary to consult the world. Quite where the demarcation lies is much disputed, but most would agree that “ $2+2=4$ ” is *a priori*, and “water is H_2O ” is *a posteriori*. Quine denies the reality of the distinction

7 ACQUAINTANCE VERSUS DESCRIPTION

As used by Russell, we are acquainted with only those things that we directly experience. Something is “known by description” if we know it only as the bearer of certain specified properties, for example, by bearing the property of being the shortest person in Russia.

8 ANOMALOUS MONISM

The doctrine in the philosophy of mind, associated with Davidson, that although mental events are identical with physical events, mental events cannot be described in ways that enable them to be brought within the framework of precise law-like generalizations (they are thus anomalous, or *unlawlike*).

9 AXIOM

A proposition taken as basic, at least for certain purposes. In the context of contemporary (post-Fregean) logic, an axiomatic system comprises various axioms, together with *rules of inference*, and the *theorems* of the system are what can be derived from the axioms using the specified rules of inference.

10 BICONDITIONAL

A statement of the form “ $p \leftrightarrow q$,” or “ p if and only if q ,” often abbreviated to “ p iff q .” A biconditional is logically equivalent to a statement of the form “if p then q and if q then p ” or $(p \rightarrow q) \wedge (q \rightarrow p)$.

11 CANBERRA PLAN

An approach to conceptual analysis and metaphysics inspired by David Lewis and enthusiastically advocated by (some) philosophers at the Australian National University in Canberra. Orthodox conceptual analysis—in the form of specifying nontrivial and noncircular necessary and sufficient conditions—has proven to be problematic: all the proposed analyses of (say) the concept of knowledge (or causation, or free will) quickly turn out to be vulnerable to counterexamples. Also, since (in the eyes of some) it is not obvious how or why analyses of our concepts will reveal anything about reality itself, it is not obvious that this mode of analysis will be of use to metaphysicians.

The Canberra Plan offers a way forward. Suppose our target concept is *causation*. We begin by assembling a collection of “platitudes” about causation, that is, claims concerning causation that ordinary people would find entirely uncontroversial. Armed with this “folk theory” of causation—which will typically be much looser than a formal analysis, and so less vulnerable to counterexamples—we then turn to the real world and the empirical sciences, with a view to isolating that feature of the world that can plausibly be viewed as playing the “role” of causation (e.g. energy transfers). The Canberra approach has its advantages, but those who look to metaphysics to provide the—perhaps radically revisionary—truth about reality will find its reliance on conventional wisdom unpalatable, as will those who share Quine’s rejection of the *a priori*.

12 CARDINAL VERSUS ORDINAL NUMBERS

In set theory the “cardinal” numbers are used to measure the size of sets. If two sets have the same cardinality, they are the same size, and they are the same size if their members can be put into a one-to-one correspondence with each other (i.e. each member in the first can be paired with exactly one member of the second, and vice versa). For finite sets, ordinary natural numbers serve perfectly well—the set $\{X, Y, Z\}$ has three members, and so is larger than $\{X, Y\}$, which has two. But Cantor established that there are different sizes of infinity: the sequence 1, 2, 3, . . . is *countably* infinite, but there are (an infinite series) of larger *uncountable* infinities. Given this, we obviously cannot

use ordinary finite numbers to register the size of these infinite sets, hence Cantor's introduction of "aleph numbers": $\aleph_0, \aleph_1, \aleph_2, \aleph_3 \dots$ which allow us to do so. The ordinal numbers were used by Cantor to classify sets possessing certain order structures, and an ordinal is defined by the ordinals that precede it. While there is just one countably infinite cardinal, \aleph_0 , there are uncountably many countable ordinals, for example, the first infinite ordinal is ω but there are also $\omega + 1, \omega + 2, \omega^2 \dots$. In Cantor's notation, ω_1 is the first uncountable ordinal.

13 CATEGORY MISTAKE

As Ryle used the term in *The Concept of Mind*, this sort of mistake involves treating a thing that belongs to one kind as though it belong to a different kind. A visitor to a University who sees the library, colleges, sports facilities, administration buildings but who then asks "But where is the University" is guilty of a category error; similarly—or so says Ryle—someone who conceives of the mind as a special kind of *thing*, rather than a collection of dispositions.

14 CLASS

Although often used as a synonym for *set*—a collection of objects construed as an object in its own right—the terms are distinguished in some leading mathematical set theories. In von Neumann/Bernays theory, there are two kinds of entities, sets and classes; moreover, there is a *proper class* (which is not itself a set) that contains all the sets as members. In Zermelo-Frankel set theory, there are only sets, and there is nothing that contains all the sets as members.

15 COGNITIVISM VERSUS NONCOGNITIVISM

The terms refer to a dispute in meta-ethics. A statement such as "Cats have four legs," made under normal circumstances, would naturally be construed as expressing a belief, and can be assessed in terms of truth and falsity—as it happens, the statement is true. Cognitivists take the view that moral statements such as "Torture is wrong" typically express beliefs, and can be assessed in terms of truth and falsity. Noncognitivists deny this, and hold that moral judgments are (say) expressions of emotional states, or desires—and hence not expressions of beliefs about factual states of affairs. As a consequence, moral statements are typically *not* capable of being true or false—they are not "truth-apt." Influential versions of noncognitivism include Ayer's "emotivism,"

Hare's "prescriptivism," Blackburn's "quasi-realism," and Gibbard's "norm-expressivism." Very different versions of cognitivism have been defended by Moore, Harman, Railton, McDowell, and Parfit.

16 CONCEPT AND OBJECT

A distinction due to Frege, who took predicate expressions to refer to *concepts*, and viewed the latter as incomplete or "unsaturated" entities, and as such distinct from *objects*, which are complete or "saturated." On Frege's view, predicates such as "... is a horse" are in need of completion in the same way mathematical functions such as " $2x$ " or " $\log. \dots$ " Since Frege held that concepts cannot be referred to by subject terms (but only by predicates) he cannot avoid the paradoxical-seeming conclusion that the concept of a horse is not a concept.

17 CONCEPTUAL ANALYSIS

See *Analysis*.

18 CONDITIONAL

A statement of the form "if p , then q " (the " p " is commonly called the *antecedent*, the " q " the *consequent* of the conditional). Logicians have distinguished a number of different varieties or "strengths" of conditional. The conditional found in standard propositional logic is the *material conditional* (in symbols " $p \rightarrow q$ "), which is only false if p is true and q false. Stronger forms of conditional attempt to accommodate the notion that the truth of p in some way necessitates the truth of q .

19 CONJUNCTION

A statement of the form " p and q ," which is true if and only if both p and q are true. In logical systems a variety of symbols for conjunction are employed, for example, " $p \& q$," " $p.q$," " $p \wedge q$."

20 CONSCIOUSNESS

What we lose when we fall into dreamless sleep, what we regain when we waken. "Conscious states" are *experiences*; the character (subjective or

phenomenal) of an experience is determined by what it is like to have or undergo an experience of that type. Quite how to accommodate consciousness within a naturalistically oriented account of the mind has been the most contentious issue in analytic philosophy of mind. The “hard problem” of consciousness is understanding how physical processes in the brain manage to give rise to experience.

21 CONTENT

What is expressed or conveyed by an utterance; the sense or meaning possessed by a proposition or mental state.

22 CONTEXTUALISM

A family of related doctrines that emphasize the essential role of context on content. An epistemological contextualist holds that what counts as “knowing” or “being justified” can and does vary from context to context (e.g. school playground, TV quiz show, scientific laboratory). In an analogous fashion, the moral contextualist holds that the truth values of attributions of moral properties vary with the context of the attributor or subject.

23 CONTRARY-TO-FACT CONDITIONAL

A sentence of the form “if such and such *had* been the case, then such and such *would* have occurred” — also known as “counterfactuals.”

24 CONVENTIONALISM

The claim, relative to a given topic, that something that is commonly thought to be wholly objective is in reality a product of human conventions (or decisions, or policies), in the manner of the conventions of etiquette, or the rules of law. Poincaré held that the geometry of space was a conventional matter; if it is, then the claim that the space of our universe is Euclidean (or non-Euclidean) is not made true (or false) by the nature of space itself.

25 COUNTERPART THEORY

The view, associated with David Lewis, that individual things are “world-bound,” that is, are able to exist in just one possible world. When we make

claims— for example, “I could have had green eyes” — which appear to involve an object featuring in other possible worlds, Lewis holds that these should be interpreted as involving *counterparts* of this-worldly objects, that is, objects that resemble objects in this world, but are numerically distinct from them. Also see *Modal Realism*.

26 CRITERION

A technical term in the philosophy of the later Wittgenstein. A criterion is evidence for the existence of what it is a criterion of, but of a (quasi-) logical—not merely inductive—kind. In the right circumstances, the presence of a criterion for X is logically sufficient for the presence of X. Wittgenstein held that we can have knowledge of other minds because there are publically accessible criteria (e.g. wincing) for the occurrence of mental states (e.g. pain).

27 DEDUCTION

An instance of reasoning in which the conclusion is established on the basis of a set of premises. In cases where the conclusion follows logically from the premises, it is impossible for the premises to be true and the conclusion false.

28 DEFLATIONARY THEORIES OF TRUTH

The view that there is nothing philosophically interesting or significant to say about truth in general. With regard to specific instances of true sentences we *can* say what their truth consists in, but only in trivial ways: “water is wet” is true if water is wet. This view—or something similar to it—is also known as the “redundancy theory of truth,” or the “disquotational theory.”

29 DENOTATION

The denotation of a singular term (such as a proper name) is the object to which it refers; the denotation of a predicate is the collection of objects that fall under it.

30 DENOTING PHRASE

The expression used by Russell to refer to a family of referring expressions, such as “any table,” “all tables,” “some tables,” “a table,” or “the table.” In *The*

Principles of Mathematics he holds that denoting phrases considered in isolation have their own meanings: “denoting concepts.” He comes to reject this view in “On Denoting,” where he argues that denoting phrases are “incomplete symbols,” which can be paraphrased out of existence without loss of meaning.

31 DE RE/DE DICTO

The distinction applies to different ways of interpreting sentences, as referring to an individual object, or not doing so. Quine provides the example “I want a sloop”: if so, there may be one particular sloop that you want, or alternatively, you may be entirely unconcerned about which particular sloop that you receive, all you want is *some sloop or other*. As one might expect, knowledge *de re* is knowledge of some particular object.

32 DESCRIPTIONS (Theory of)

Russell’s treatment of descriptive phrases in “On Denoting.” A statement such as “The present king of France is bald” is taken to be equivalent to a statement that asserts that there exists a unique object that is both the present king of France and is also bald.

33 DISJUNCTION

A sentence of the form “*p* or *q*” (in logical symbols $p \vee q$). There are two forms of disjunction: the *inclusive* and the *exclusive*. In the latter case, “*p* or *q*” is true *iff* one or other disjunct is true, but not if both are true. In the former case, “*p* or *q*” is true *iff* *p* either or both of the disjuncts are true. Standard logic texts assume that most of the disjunctions encountered in ordinary English are of the inclusive variety.

34 EMOTIVISM (Expressivism)

A position in meta-ethics according to which ethical statements do not express propositions, but rather are expressive of emotional attitudes. Influential variants of this form of *noncognitivism* were defended by A. J. Ayer and C. L. Stevenson.

35 ESSENTIALISM

The doctrine that (at least some) entities possess (at least some) of their properties essentially, that is, that they could not exist without those properties.

36 ETHICS VERSUS META-ETHICS

We can inquire into what is right and wrong, we can also inquire into the *nature* of right and wrong. If a given course of action is the morally right one, what is it that we are saying (or doing) when we say as much? The investigation into what is actually right and wrong (in given circumstances) is “first-order” or “normative” ethics. The inquiry into the nature of ethics is “second-order” or “meta-ethics” (or “meta-normative”). Among the main issues in meta-ethics are: is morality objective, or subjective? Is morality culture-relative? Do moral facts and moral truths exist? If so, in what do they consist, and where do they originate? How does morality influence our actions and motivate us?

37 EXTENSION VERSUS INTENSION

A predicate’s extension is the sum total of all the objects to which it applies—hence the extension of “round” is the class of round things. The intension of a predicate is the condition that must be met for the predicate to apply. As Quine noted, the predicates “creature with a heart” and “creature with a kidney” have the same extension, but they are associated with different intensions. These terms are also sometimes used in connection with singular referring expressions, for example, “Superman” and “Clark Kent” are names with different intensions, but the same extension. In a so-called extensional context it is possible to substitute coreferring *salva veritate* (i.e. without affecting the truth value of the sentence), whereas in an “intensional context” this is not possible. Hence, “Superman has two hands” is extensional, whereas “Lois Lane believes Superman is superstrong” is not: there is no guarantee that she believes that *Clark Kent* is superstrong.

38 EXTERNAL VERSUS INTERNAL RELATIONS

See *Relations*.

39 EXTERNALISM VERSUS INTERNALISM

Externalism in the philosophy of mind and language is the doctrine that the contents of (at least some kinds) mental state and (certain classes of) statement are determined by, and dependent on, features of the world that are external to the mind of the relevant subject. Internalists deny that the states or statements in question are world-dependent in this way, and thus uphold a Cartesian view of the autonomy of the relevant aspect of the mental with regard to the rest of reality. Externalists in epistemology hold that being suitably related to certain aspects of the external world can be enough for genuine knowledge, even when subjects are not aware of the relevant relationships. Also see *Wide and Narrow Content*.

40 FACT

There are those (e.g. the early Wittgenstein) who hold that facts are a distinctive kind of entity in their own right, composed (say) of one or more objects, their relations and properties. Others adopt a more minimalist, deflationary stance. Since there is little or no difference between the meanings of “It is a *fact* that *P*” and “It is *true* that *P*,” it is not clear that we need to posit facts in addition to truths.

41 FICTIONALISM

Sherlock Holmes does not exist, yet is it not true that Sherlock Holmes lived at 221b Baker Street? The obvious way forward is to say that it is not really or straightforwardly true that Holmes lived at 221b, but it is *true in Conan Doyle’s fiction* that he lived at that address. When we make claims such as “Holmes lived at 221b” the qualifying *in the fiction* is not explicit, but it is implicit. With a view to securing ontological savings, the fictionalist about a given area of discourse argues that claims made in this area are best construed as analogous to claims about fictional characters. This allows us to continue to talk *as if* the relevant entities exist—which is often very useful—without actually believing that they do. In recent decades, Harty Field (1980) has defended fictionalism about mathematics, Rosen (1990) recommends a fictionalist stance about possible worlds, others have defended fictionalisms about ordinary material objects, moral claims, and scientific theories. Precisely what fictionalism with regard to a given area of discourse amounts to depends on the accompanying analysis of construing fictional discourse itself, which remains a contentious issue.

42 FIRST-ORDER LOGIC (LANGUAGE)

Formal systems equipped with quantifiers able to range over objects but *not* properties are said to be first-order. In second- and higher-order logics, the quantifiers are not confined to objects, and so can take properties as values. A typical first-order quantifier is “for every object in the universe of discourse. . . .” A typical second-order quantifier is “for every property in the universe of discourse. . . .” The latter quantifier is equivalent to “for every *set* of objects . . .” if properties are identified with sets, as they are in extensional logics. In *Philosophy of Logic* Quine claimed that second-order logic is “set theory in sheep’s clothing,” a controversial claim that logicians continue to debate.

43 FORMAL MODE VERSUS MATERIAL MODE

A distinction employed by Carnap, particularly in *The Logical Syntax of Language*. In the material mode, words referring to things are construed in the usual way (as referring to things), in the formal mode, we switch focus to the words themselves, and make claims about them as such. For example, the claim that “a dog is an animal” is in the material mode, whereas “‘dog’ is a word for a type of animal” is in the formal mode.

44 FORMALISM

In the philosophy of mathematics, formalism is the doctrine that mathematicians are not exploring the contents of an abstract Platonic realm, but rather the consequences of various rules for manipulating symbols. On this view, mathematics is akin to a game, and its terms are construed as nonreferring. Ethical formalism is the view that morality derives from high-level abstract principles that can be applied universally.

45 FOUR-DIMENSIONALISM

The theory of time according to which past present and future are equally real, they are just located at different points in time, on analogy with occupying different positions in space. This solves certain problems associated with *presentism* but makes it difficult to accommodate the apparent flow of time, which finds expression in the “moving ‘now.’” See also *Presentism, Growing Block*, and *Time: A-series and B-series*.

46 FUNCTION

In logic and mathematics a function can be thought of as a mapping that connects a class of inputs with a class of outputs, so that each input is mapped onto just one output. In a more formal vein, the input to a function is the *argument*, and the output the *value*. In the usual symbols, a function f takes a argument (x) and generates the value $f(x)$. The class of potential inputs is called the *domain* of the function, and the class of potential outputs the *range* of the function.

47 FUNCTIONALISM

Functionalism, a doctrine in the philosophy of mind that has been influential in recent decades. According to the functionalist, roughly speaking, the mental character of a mental state is determined by the casual role of the state in a complex system, where the latter mediates between perception and action. The functionalist approach constitutes an advance over behaviorism in two crucial respects. First, it regards mental states as internal, whereas behaviorism tended to be a “blackbox” theory, disavowing interest in internal processes. Second, it is concerned with the functional role of these states in influencing *each other*, as well as in their role in producing behavior and being responsive to perceptual “inputs.” The theory differs from the classical “identity theory of the mind” by allowing that the same kind of state can be realized in different hardware in different creatures or at different times: that is, like functional states in general, mental states are *multiply realizable*. Nevertheless, an internal state counts as mental because of the contribution it ultimately makes to behavior, just as a computational state counts as part of a program (and not just an incidental part of the hardware) because of its role in producing the results for which the program is designed.

48 GRAMMAR

In linguistics, the term refers to the rules according to which sentences can be constructed from smaller elements of language. Those who believe there is a *universal grammar* hold that there are categories, structures and rules that are common to all human language. “Grammar” is also a quasi-technical term in Wittgenstein’s later philosophy, referring to the rules (in language-games) that determine what it makes sense to say, and what it *does not* make sense to say. For Wittgenstein,

Grammar is not accountable to any reality. It is grammatical rules that determine meaning (constitute it) and so they themselves are not

answerable to any meaning and to that extent are arbitrary . . . The rules of grammar cannot be justified by showing that their application makes a representation agree with reality. (*Philosophical Grammar*, pp. 184–6)

Quite what Wittgenstein meant by saying that grammar is arbitrary and autonomous has been much debated.

49 GÖDEL'S THEOREMS

Informally, Gödel's first incompleteness theorem says that for any formal system that is sufficiently powerful to express arithmetic, there exist sentences that are true but that cannot be proved. His second theorem states that systems of this kind cannot prove their own consistency. Gödel's results have implications for some philosophies of mathematics—they are generally assumed to doom Hilbert's program of finding a complete and consistent set of axioms for all of mathematics—but there is less consensus with regard to their broader implications (e.g. for artificial intelligence).

50 GROWING BLOCK

The theory of time according to which the past and the present are real, but the future is not, thus reality grows in the direction of the future and the present is its latest outermost surface. On this view the passage of time consists in the creation of new "slices" of reality, moment by moment. See also *Presentism*, *Four-dimensionalism*, and *Time: A-series and B-series*.

51 HIGHER-ORDER LOGIC

See *First-order Logic*.

52 IDENTITY

The relation every object bears to itself and itself alone. If a and b are identical (if $a = b$), then " a " and " b " are simply names of one and the same thing. Leibniz formulated two principles governing identity. According to the principle of the *indiscernibility of identicals*, if $a = b$, then a and b have precisely the same properties. Leibniz's principle of the *identity of indiscernibles* states that if a and b have exactly the same properties, then $a = b$. Whereas the former is uncontroversial, the latter is not.

53 IFF

A common abbreviation for “if and only if”; also see *Biconditional*.

54 INTERPRETATION

In logic an interpretation is not a (sometimes) contentious reading of a literary text, but something more precise: an assignment of meanings (or semantic values) to the elements of the system. In simple formal language, we will have logical and nonlogical symbols, and an interpretation of the system will specify precise meanings for each symbol. In the case of the predicate calculus, the logical symbols may include quantifiers such as $\forall x$ and $\exists x$, along with truth functional connectives such as “&,” “ \vee ,” and “ \sim ,” and the interpretation will specify how formula containing these symbols are to be evaluated. The nonlogical symbols will include (for example) predicate letters, $F, G, H \dots$ and names, $a, b, c, d \dots$. The interpretation will specify a domain (of discourse), and for each predicate letter will indicate the collections of objects to which they apply, and for each name the object it refers to.

55 INTENSION/INTENSIONALITY

See *Extension*.

56 INTENTIONALITY

The capacity of minds (or mental states) to be directed at, be about—or represent—things or events or properties. Brentano took intentionality to be the distinguishing feature (or “mark”) of mental phenomena. When used in this sense it should not be conflated with “intention” in the sense of “having an aim or objective”—intentions in the latter sense are just one of the many kinds of mental state that can have intentionality, for example, beliefs, desires, hopes, fears, hates, and likes. Intentionality should also not be confused with *intensionality*, which is a property of sentences or referring expressions.

57 INTUITION

The immediate (noninferential) apprehension of the truth of a proposition, or the direct awareness of a proposition or concept. In the Kantian tradition, intuitions are passive sensory representations, and thus the raw material of

empirical knowledge. As most commonly used in contemporary analytic philosophy, intuitions are judgments as to the appropriateness to a given scenario of a description or a concept, and hence routinely appealed to in conceptual analysis. Since intuitions begin where arguments end, or so it is sometimes said, they are regarded by some as suspect.

58 INTUITIONISM

In ethics, the view—championed by G. E. Moore in *Principia Ethica*—that moral truths are objective, unanalyzable, and known to be true by a special faculty of intuition. In the philosophy of mathematics, the view—championed by Brouwer—that a mathematical statement is true only if a proof for it can be given, and that the mathematics is the exploration and development of mental constructions. Intuitionists are opposed to logicism and Platonism. Intuitionistic logic rejects the law of the excluded middle (p or not- p).

59 JUDGMENT

Broadly speaking the (mental) affirmation that such and such is the case. A theory of judgment is typically a theory both of what it is to believe something, and *what it is* that is judged (e.g. a proposition) when something is believed. Although very different accounts of judgment have been defended, the underlying problem is the mind's ability to represent actual, or merely possible, states of affairs. The term is not much used—in this sense—in modern-day analytic philosophy.

60 LOGIC

The systematic study of inference. Since inference comes in different forms, there are different branches of logic. In *inductive* reasoning, the premises of an argument lend support to the conclusion, but their truth does not conclusively establish the truth of the conclusion, for example, "We have randomly sampled 5000 swans, they have all been white, and this confirms the hypothesis that all swans are white." Inductive logics typically represent inductive inferences in terms of probability, and the latter concept has itself been intensively studied by analytic philosophers. In *deductive* logic, the premises of an argument entail the conclusion, that is, it is impossible for the conclusion to be false if all the premises of the argument are true—or to put it another way, the truth of the premises guarantees the truth of the conclusion. Deductive arguments with this property are said to be valid. An argument is *sound* if, in

addition to being valid, its premises are actually true. Since deductive validity is a property arguments possess in virtue of their *general form*, deductive logic is a study of valid argument forms. Modern (deductive) logic was born when Frege invented the propositional and predicate calculi in the nineteenth century, but that was just the start: there are now modal logics, tense logics, deontic and epistemic logics, all with their distinctive features and applications, though generally these are based on—or consist of augmentations to—classical Fregean logic.

61 LOGICAL ATOMISM

A doctrine common to Russell and Wittgenstein (in the period of *Philosophy of Logical Atomism* and the *Tractatus*) according to which ordinary language statements can be analyzed down to primitive expressions, the referents of which are the elementary constituents of reality.

62 LOGICAL CONSTRUCTION

In “Logical Atomism” Russell puts forward this methodological recommendation: “Wherever possible, substitute constructions out of known entities for inferences to unknown entities.” The definition of numbers in terms of equinumerous classes was the earliest application of the maxim, but there were several others. Russell came to view propositional functions as logical constructions from ordinary propositions, and he later extended the approach to the physical world, arguing (at one point) that material objects were logical constructions of sense-data. Details aside, the idea that we can reduce our ontological commitments by logical analysis and paraphrase was hugely influential in analytic philosophy.

63 LOGICAL SYMBOLS

63.1 Propositional Connectives

These operate on whole propositions (or sentences), which are symbolized by upper case letters, for example, the proposition “The cat is on the table” could be symbolized by “P,” and “Paris is the capital of France” by “Q.” The standard propositional connectives can each be symbolized in various ways, as follows.

Negation: with P interpreted as above, the negation of P is “It’s not the case that the cat is on the table,” which can be symbolized thus: $\neg P$, $\sim P$, $\neg P$.

Conjunction: the logician's term for "and," which is commonly symbolized thus: $P \& Q$, $P \cdot Q$, $P \wedge Q$, PQ . Hence "The cat is on the table and Paris is the capital of France" could be symbolized by " $P \& Q$." A conjunction is true only if both conjuncts are true.

Disjunction: the logician's term for "or." The disjunctive statement "The cat is on the table or Paris is the capital of France" would most commonly be symbolized thus " $P \vee Q$." Such statements are true only if one or other *or both* of the component disjuncts are true.

Material Conditional: the logician's term for statements of the form "IF . . . THEN . . .," typically symbolized " $P \rightarrow Q$," " $P \supset Q$," so the latter might mean "If the cat is on the table, then Paris is the capital of France," which is taken to be *false* only when the antecedent (P) is true, and the conclusion (Q) is false.

Biconditional: statements joined by "if and only if," symbolized by \leftrightarrow or \equiv , so " $P \leftrightarrow Q$ " might mean "The cat is on the table if and only if Paris is the capital of France," which will be true only if P and Q both have the same truth value.

63.2 Quantificational Logic

In this context " Fx " is taken to mean " x is F ," so if " F " means ". . . is a cat" and " G " means ". . . is happy," then Fx states that x is a cat, and Gx states that x is happy. Expressions such as Lxy assert that the relationship L exists between x and y , hence if L is interpreted to mean *loves*, then Lxy states that x loves y . An important class of arguments make claims about *all* F 's and *some* F 's. In the terminology of logicians, "all" and "some" are "quantifiers." By using quantifiers we can make claims such as "All F 's are G 's," which can be symbolized thus:

$$\forall x(Fx \rightarrow Gx) \text{ or equivalently } (x)(Fx \rightarrow Gx)$$

This formula can be read thus: "for all x , if x is F , then x is G ." The claim that "some F is G " is symbolized thus:

$$\exists x(Fx \& Gx)$$

Or "There is at least one x such that x is F and x is G ." Quantifiers can usefully be deployed in disambiguating ordinary language sentences. Consider "Everyone loves someone." Does this mean there is some particular person whom everyone loves, or that everyone has someone (not necessarily the same someone) whom they love? Since the sentence in question can have both

meanings, it is ambiguous. These different claims expressed without ambiguity in the language of quantificational language.

$$\exists x \forall y (Lyx)$$

If we restrict the “domain” the quantifiers operate over to people, this formula says “there is some person x who is such that everyone loves x .” We can express the other meaning in this way:

$$\forall x \exists y (Lxy)$$

This formula says “For any person x , there is some person y that x loves,” or more colloquially, “Everyone has someone they love.”

64 LOGICAL TRUTH

A statement that is true by virtue of logic alone. Such a statement may be a theorem of a particular logical system, or true under any *interpretation* of the system.

65 LOGICALLY PROPER NAME

According to Russell’s theory of descriptions, many expressions that have the form of referring expressions, are seen not to be such when properly analyzed. The expressions that *do* refer, even on the final analysis, are the logically proper names. Russell (at one phase) held that the only logically proper names are words such as “this” or “that” when used to refer to items we directly apprehend, for example, sense-data.

66 LOGICISM

The doctrine in the philosophy of mathematics, according to which mathematics is reducible to logic.

67 MANY-VALUED LOGIC

A logical system in which there are more than two truth values. In some of these nonclassical systems there are three truth values (true, false, indeterminate), in others there are four, in others there are an infinite number of

gradations between being straightforwardly true, and straightforwardly false. These systems have intrinsic interest for logicians and mathematicians, but there are also philosophical motivations, for example, accommodating the phenomenon of *vagueness*, the “openness of the future” (on some views, statements about future contingencies are neither true nor false), or the various logical *paradoxes*.

68 MATERIAL IMPLICATION

See *Conditional*.

69 MATERIAL MODE

See *Formal Mode*.

70 MEANING

Some sounds and inscriptions possess meaning, others do not. A central philosophical issue pertaining to meaning is arriving at an understanding of the facts that go to determine the meanings that meaningful expressions have. On some views, the meanings of expressions in a public language ultimately flow from the contents of the mental states of the speakers of the language (e.g. Grice’s program of analyzing expression-meaning in terms of speakers’ meaning, and reducing speakers’ meaning to speakers’ intentions). Others deny that linguistic representation can be analyzed in terms of mental representations, and look instead to truth maximization (e.g. Davidson) or systems of public rules (e.g. those seeking to systematize the insights of the later Wittgenstein). So-called semantic theories of meaning focus on the theories that systematically assign meanings to the words and sentences of a language. The distinctively philosophical issue here is the general *form* such a semantic theory should take. Influential theories along these lines include Carnap’s *possible world semantics*, and Davidson’s *truth-conditional semantics*, together with their various descendants.

71 MEREOLOGY

From the Greek for “part,” *meros*, it is the study of the formal relations of parts and wholes. It is extensional, in the following sense: a mereological sum is a collection of parts and it is a different sum if any of the parts are different. So you become a different mereological sum when you lose an atom from your

body. In this way, it is quite different from our ordinary notion of *object* or *substance*. It is equivalent to Locke's notion of *mass of matter*. *Mereological nihilism* is the view that collections of parts do not constitute real objects: there are, for example, no tables, only atoms/parts arranged table-wise. *Mereological universalism*, on the other hand, is the view that any and every collection of parts constitutes an object. So the little finger of your left hand and the dark side of the moon constitute an object. It is just that only some objects are of any interest—for example, have any explanatory value.

72 MODALITY

There are different ways (or modes) in which sentences or statements can be true. A statement is *necessarily* true if it could not be false; a statement is only *contingently* true if it is not necessarily true. Tense is another important mode: some statements are true at all times, others only in the present, others only in the past or the future.

73 MODAL LOGIC

The expression usually refers to the logic of possibility and necessity, that is, the deductive behavior of “it is possible that . . .” and “it is necessary that. . .” The dominant contemporary approaches derive from C. I. Lewis, in publications dating back to 1912, who added two modal operators— \Box , meaning *it is necessary that*, and \Diamond , meaning *it is possible that*—to the propositional and predicate calculi. There are many different systems of modal logic, two of the better-known are *S4*, which includes the axiom $\Box A \rightarrow \Box \Box A$ (“if *A* is necessary, then it is necessary that *A* is necessary”), and *S5* which also includes $\Diamond A \rightarrow \Box \Diamond A$ (“if *A* is possible, then necessarily *A* is possible”). The now-standard possible world semantics (or model theories) of modal logic were supplied by Kripke and others. In Kripke's (Leibniz-inspired) approach, a proposition is necessarily true if it is true in all possible worlds, whereas a proposition is contingently true if it is true in only some possible worlds.

74 MODAL REALISM

The doctrine, defended by David Lewis in *The Plurality of Worlds* (1986) that other possible worlds are just as real as this world. For Lewis, a “world” is single spatio-temporal system containing one or more concrete material objects, the cosmos as a whole consists of an infinite number of disconnected systems of this kind; everything that *could* happen *does* happen in one (or

more) of these worlds. Modal realism is not an economical theory—indeed, it may well be that no theory is more vulnerable to Occam’s razor—and critics further allege that it fails to do justice to the difference between the actual and the merely possible. Lewis argues that it is superior to all other accounts of possible worlds, and has the additional benefits of providing highly economical solutions to many other central problems of philosophy. For example, we no longer need a special category of modal facts to make sense of modal discourse—“pigs could have been able to fly” is made true by actual pigs with this ability in a different part of the cosmos. If we follow Lewis and identify propositions with sets of worlds—for example, the proposition “pigs can fly” with the set of worlds containing pigs that can fly—we need not posit propositions as a special category of entity over and above other kinds of thing. We can pursue a similar policy with *properties*—for example, taking *green* to be the set of all green things—and so eliminate properties from our ontology too. *Ersatz* modal realists, as they are frequently called, take possible world talk seriously, but reject modal realism. Instead they posit a variety of entities—sets of propositions, sets of sentences, sets of states of affairs, for example—to do the job of Lewis’ *real* possible worlds. See also *Counterpart Theory*.

75 MODEL THEORY

A model is an *interpretation* of a formal system, and model theory is the general study of interpretations (in this sense of the term). Many important logical properties can usefully be characterized in model-theoretic terms. For example, if *P* and *Q* are sentences in a formal system *S*, then *Q* is a *consequence* of *P* if and only if *Q* is true in all models in which *P* is true; *P* is *valid* in *S* if and only if *P* is true in all models of *S*.

76 MODUS PONENS

The name for the valid rule of inference that, given some proposition *P* and a proposition of the form if *P* then *Q*, *Q* can be derived.

77 MODUS TOLLENS

The valid rule of inference that from a proposition of the form if *P* then *Q*, and not-*Q*, then not-*P* follows.

78 NAMES

Names designate individuals, but not via some characteristic, as would a predicate. J. S. Mill said that names denote but do not connote, that is, do not have a sense or meaning of the sort one might hope to look up in a dictionary. Philosophers within the analytic tradition have disagreed about how names manage to latch on to the thing they name, given that it is not explicitly via a property. Frege, for example, thought that a name must be short-hand for some definite description, so “Aristotle,” for example, might mean, for a given speaker, “the author of the *Metaphysics*.” Those who do not like this so-called *descriptive theory of names*, initially preferred an *acquaintance theory*, according to which direct denotation depends on direct confrontation with an object. This led Russell to believe that we can only name objects in our immediate experience—sense-data (See *Logically Proper Names*). Later philosophers extended acquaintance to objects in the external world, and, following Kripke, tend to go for a *causal theory of names*, according to which objects are dubbed with names at a kind of baptism, and subsequent use depends on a causal chain reaching back to this event.

79 NECESSARY AND SUFFICIENT CONDITIONS

P is a necessary condition for Q if Q cannot be the case unless P is the case. P is a sufficient condition for Q if P’s being the case guarantees that Q is the case. So *being human* is a necessary condition for being an Englishman, for nothing can be an Englishman without being human. But it is not sufficient for something can be human and not be an Englishman—for example, it might be a Frenchman or an English woman. But being an Englishman is sufficient for being human, for one cannot be an Englishman without being a human being. In general, if P implies Q, then P is sufficient for Q and Q is necessary for P.

80 NEGATION

The negation of a proposition is its denial or contradictory. The negation of P is not-P.

81 NOMINALISM

Nominalism is the contrasting position with realism in the theory of universals. There is generality in the world as we understand it: many things are red, square, human, etc. The question for the “theory of universals” is where this

generality comes from. The realist says that it exists out there in the world itself, quite independently of how we classify things: there are universal features—properties—out there as real as the objects that possess them. The nominalist says that such general features are the creatures of our activity of categorizing—“naming”—and that nothing general exists in the world itself.

This issue dominated much of classical and mediaeval metaphysics, and the precise definition of both nominalism and realism through the history of philosophy is much more difficult than the seemingly simple characterization above suggests.

82 NORMATIVITY

Philosophers have been preoccupied at least since David Hume with the fact-value distinction, but thought of values as consisting of such things as ethical and possibly aesthetic judgments. More recently philosophers have become aware that classifying things as correct or incorrect, rational or irrational also has a character essentially different from a simple fact. Natural—paradigmatically physical—processes simply happen, usually according to natural laws; they are not, as such, correct or incorrect, reasonable or unreasonable. But most of our mental activity, not just our moral or aesthetic judgments are evaluated for standards of right or wrong: assessment of evidence, accurate representation of facts, arithmetical calculations, etc. This pervasiveness of normativity has been judged to present a challenge to physicalism and to lead one to a naturalism more broadly conceived, if such a thing is possible.

83 OBJECT LANGUAGE

An object language is one that speaks about things in the world outside of the language itself. It contrasts with a *meta-language* that talks about the language that talks about the nonlinguistic objects.

84 ORDINARY LANGUAGE

This is the language of everyday, nonscientific, nontechnical, nonphilosophical discourse. In the phrase “ordinary language philosophy” it signifies the philosophical tendency, originating in postwar Oxford and associated with J. L. Austin, which tended to think that classical philosophical problems arose because philosophers theorized too swiftly on the basis of certain linguistic facts, ignoring subtleties of discourse. So, for example, ordinary language philosophers thought that the sense-datum theory of perception arose because it

was assumed that, if an object looks red there must be something red of which one is aware, ignoring what one can achieve by deploying “looks,” “seems,” “appears” idioms, etc. It took up the deflationary attitude to metaphysics characteristic of logical positivism, but did not base this on verificationism, formal logic, and natural science, but on the belief that ordinary language was so rich that, if one were careful, one could characterize situations without resorting to philosophical theories. Ordinary language philosophy also makes use of the *paradigm case argument*.

85 PARADIGM CASE ARGUMENT

Almost all philosophical terms, or their close neighbors, have uses in ordinary nonphilosophical contexts. We talk of people acting freely, of coming to know things, perceiving things in the world, etc. And it is plausible to claim that we learn these terms from the standard cases to which they are applied. Free choice is when someone, under no outside pressure, decides to do what he wants: seeing a physical object is when someone reports that he can see the orange we are holding up in front of him: knowledge is when they can get ten out of ten in a question session. The paradigm case argument asserts that, as these terms are defined by reference to these standard circumstances, it makes no sense to be skeptical about whether there are such things as free choice, knowledge, perception of the world, etc.: the paradigm cases exist, so must the phenomena of which they are paradigms, for that is how the terms are defined.

The argument rests on one of the major assumptions of ordinary language philosophy, namely that our ordinary discourse is free of implicit assumptions or theories. If that were so, then freedom, or seeing or knowing (for example) must be exemplified by the cases paradigmatically used to illustrate them. But if there are implicit assumptions, they could turn out to be wrong.

86 PARADOX

A paradox arises when two contradictory propositions can be derived from seemingly acceptable premises and forms of argument. There seem to be at least two types of philosophically interesting paradox. In one, a seemingly obvious truth is countered by a clever piece of reasoning. An instance of this is Zeno’s paradox that, if a tortoise is given a start over a hare, the hare could not catch up with it in a finite time. We know that a hare could easily catch a tortoise, but Zeno has an argument that seems to show that it would need an infinite number of steps to do so.

The other kind is when a seemingly straightforward concept is shown to generate a contradiction. An instance of this is the *barber paradox*. In a village

there is a barber who shaves all and only those men of the village who do not shave themselves. This seems a perfectly straightforward idea, but then ask whether the barber shaves himself. If he does, then he does not, for the barber does not shave anyone who shaves themselves. But if he does not shave himself, then he does for he shaves all those who do not shave themselves. In fact the notion of a barber who shaves all and only those who do not shave themselves is incoherent.

87 PHENOMENALISM

Phenomenalism is a radical empiricist account of the physical world, according to which physical objects are, in J. S. Mill's words, just "actual and possible sensations." That is, physical objects are reduced to the experiences that are or might be had of them. Its roots are in Berkeley's idealism—which is a theistic phenomenalism—and Hume's empiricism, followed by Mill and Russell and finally the logical positivists. These latter turned it from being a substantive ontological theory into a linguistic one. Linguistic phenomenalism holds that statements about physical objects are equivalent to statements about actual and possible observations.

88 PLATONISM

Platonism in modern philosophy (as opposed to theories contained in Plato's texts, the interpretation of which is controversial) is the theory that abstract objects exist, and do so independently of their instances. This is called *ante rem* realism, as opposed to the *in re* realism, attributed to Aristotle, according to which they exist but only in their instances. Through most of the history of philosophy, the abstract objects in question are universals—properties and general features that can belong to many things. But in modern debates numbers, propositions, and sets are also generally classed as abstract objects, and a Platonist about them will affirm their objective existence, independently of either human thought and practices or the concrete situations involving them.

89 POSSIBLE WORLD

As well as the way things actually are—which, in its totality, is termed the *actual world*—there are ways things might have been. We normally think of these latter as slight deviations from what actually is the case—I might have had an extra cup of tea this morning. But one can also think of the totality

so obtained as a “different possible world.” A possible *world*, as opposed to a fragmentary scenario, involves assigning a truth value to every proposition, so every possible state of affairs either is or is not the case.

This jargon was brought into philosophy to provide a semantics for modal statements. So “I might have had an extra cup of tea this morning” denotes those worlds in which I did have an extra cup. The ontological status of possible worlds is a matter of controversy. David Lewis thought they were concrete objects just as is the actual world, but standing in no spatio-temporal relation to the actual world. Most philosophers think of them as abstract objects, such as sets of propositions. “Possible world” jargon has proved very useful, even though most philosophers would admit that they do not know what it is they are talking about!

90 POSITIVISM

The term “positivism” was initially associated with August Comte in the mid-nineteenth century, and signified “exactness.” Positivism, therefore, was concerned to pursue the most exact and rigorous standards in science. In philosophy, the term is mainly associated with *logical positivism*, the ideology of the Vienna Circle. It combined a radical empiricism of a scientific kind, with formal logic, following Russell’s logical atomism, and denied meaningfulness to all metaphysical and traditional philosophical theories, on the ground that they could not be given exact scientific or empirical expression. “Legal positivism” denies the existence of any natural law and says, roughly, that law is just what judges decide.

91 POWER SET

A term in set theory; a power set of a given set is the set of all subsets of that set.

92 PREDICATE CALCULUS

Modern formal logic has at its foundation the propositional and predicate calculi. The former maps the logical relations of complete propositions, whereas the predicate calculus, like the syllogism, breaks propositions down into subject-predicate form, though, unlike the syllogism, it includes relations among the predicates it can handle. As well as the same logical connectives as those in the propositional calculus, it includes predicate letters, variable, names, and universal and existential quantifiers.

93 PRESENTISM

A position in the philosophy of time according to which only the present exists. The future, after all, does not yet exist and the past has ceased to exist. If one were to take the view that the present is a mere instant, then all that existed would be an instantaneous interface between two nonexistents! Hence there is a certain pressure to treat the present as having some duration. See also *Growing Block*, *Four-dimensionalism*, and *Time: A-series and B-series*.

94 PRINCIPLE OF EXCLUDED MIDDLE

This is the principle that, for any proposition P, either P or not-P: there is no third or middle option. It is closely related to the *principle of bivalence* that holds that every proposition is either true or false.

95 PRIVATE LANGUAGE

A private language—often termed a *logically* private language—is, strictly speaking, a language that, in principle, only one person can understand. In practice the term is used to signify the doctrine that certain terms have the meaning that they do for a given speaker because of the kind of private experiential episodes they refer to. For example, what I mean by “red” is determined, at least in part, by what it is like for me to experience certain things, such as tomatoes. Your experience might be quite different, hence the meaning you attach to the same word, “red,” might be quite different. This is important for empiricists and most Cartesians, because, for them, our conception of the world and the language we use to articulate it rest on private experiences of this sort.

Wittgenstein’s notorious polemic, known as the *anti-private language argument* is an attempt to show that this account of language is incoherent because, without public collusion, no fixed meanings can be established. There are various interpretations of what the argument is supposed to be that shows this.

96 PROPER NAMES

See *Names*.

97 PROPERTY

A property is whatever is signified by a predicate. A first-order property is one possessed by first-order individuals, that is not by groups, sets, collections,

etc. of objects. The issue of whether one needs to “include properties in one’s ontology” is more or less equivalent to the question of whether one is a nominalist or realist about universals. The nominalist view on properties, associated with Quine, is that to say that an object possesses a property is just to say that a predicate is true of it: there are only individuals in one’s ontology, of which certain things are true, not individuals and properties, which are what makes ascription of the predicate true.

98 PROPOSITION

A proposition is what is expressed by an indicative sentence—if it succeeds in expressing anything. Propositions are thought to be needed in addition to sentences for various reasons. For example, the English sentence “some roses are red,” the French sentence “quelques roses sont rouges,” and the Hungarian sentence “nehány rozsa piros” are different sentences, but all say the same thing. And “it is raining today” says different things on different days. Some philosophers hold that certain perfectly meaningful indicative sentences of which the subject term lacks reference—such as “the present King of France is bald”—fail to say anything and express no proposition.

99 PROPOSITIONAL ATTITUDE

Thoughts are, in general, held to have propositional content; for example, one might think *that* the weather is good today. In addition to this content there is the *attitude* that one might have toward it. You might *believe* that the weather is good, or *hope* that it is, or *fear* that it will not be, and many more. Attitudes can be held to objects as well as propositions, as one might *fear* Zeus, or *worship* him, or *love* someone.

100 PROTOCOL SENTENCE

A protocol sentence (or statement) is an observational sentence, that is one that reports the foundational evidence for any further empirical beliefs or theory. The term derives from Carnap. He originally thought that such sentences reported subjective experience of the “blue patch now” kind, but then modified his position so that physical-object concepts could figure in such reports. For an orthodox empiricist, such observational sentences would be reporting foundational facts of some kind, but for a logical positivist like Carnap it is a matter of convention what kinds of statements fit this role.

101 PSYCHOLOGISM

This was a popular theory of logic in the nineteenth century, when the laws of logic were characterized as *laws of thought*. This expression need not denote psychologism, because it might signify, in a sense correctly, that they specify how one *ought* to think, namely logically. But in the post-Kantian tradition the “laws of thought” were thought to be part of the structure of the mind, rather as are the Kantian categories. Frege led the assault on this doctrine, holding that the laws of logic are objective, Platonic entities, quite independent of human thought.

102 QUANTIFIER SHIFT FALLACY

It does not follow from the fact that every person has a father, that there is a father of everyone, nor from the fact (if it is a fact) that everything has a cause, that there is a cause of everything. To deploy arguments of this sort is to commit the quantifier shift fallacy. It has the name it does because it involves reversing the position of quantifiers in an invalid way. One is moving from “*Everything has some F*” to “*There is some F of everything*.”

103 REDUCTION

Reduction can be thought of as applying either to entities or to the theories that characterize them. When it is the entities, “*Xs reduce to Ys*” means that *Xs* are “*nothing but/nothing over and above Ys*.” So one might say that macroscopic objects are *nothing but* clouds of atoms. If one is talking about theories, then one has the sterner task of trying to show how the laws in the more basic theory make true or give rise to the laws in the theory one hopes to reduce. This latter was the objective of the positivists in trying to devise a *unified science*. Talk of reductionism in the philosophy of science is mainly concerned with the relationship of the special sciences to physics, but most talk about reductionism among philosophers concerns the philosophy of mind—whether mental states can be reduced to—shown to be nothing over and above—brain states or dispositions to behave.

104 REFERENCE

Reference is the picking out of a particular object, theory of reference is the account of how this is possible in language and thought. One way is by names that are assigned to objects, another is by demonstratives “*this*,” “*that*,” etc.

Another way is by noun phrases that seemingly pick out a particular object, such as “the present King of France.” This is where the fun begins. Russell believed that genuine referents actually entered into the propositions and thoughts in which they were referred to, and that if a referent did not exist then you could not be thinking the thought that you thought you were thinking, which Russell thought absurd. So he devised his theory of descriptions, according to which such noun phrases were *definite descriptions* and not referring expressions. The idea that one cannot refer to nonexistent things has remained orthodox among analytical philosophers until quite recently. So, when discussing propositional attitudes in the philosophy of mind, one can believe in, love, or fear Zeus, but when doing philosophy of language, one cannot refer to him.

105 REFERENTIALLY OPAQUE/TRANSPARENT

A transparent context is one in which you can substitute coreferentials and be sure of preserving truth value and an opaque context is one in which you cannot. So if it is true that Cicero denounced Cataline and that Cicero is the same person as Tully, then it must be true that Tully denounced Cataline. But it does not follow that if John believes that Cicero denounced Cataline, he must believe that Tully did, for he may not know that Cicero is Tully. Opaque contexts are also called intensional (notice the “s”) and transparent are extensional.

106 RELATIONS

A relation is a property or predicate (depending on whether one is talking *de re* or *de dicto*) that connects two or more objects. A relational predicate has two or more argument places and is *polyadic* and a monadic predicate has just one. “X is to the left of y” and “x is between y and z” are relational, but “x is green” is monadic. Relations could be expressed in logical form only with the advent of modern symbolic logic. In the Aristotelian tradition, the attempt was to express them as simple monadic predicates, which is not possible.

107 RIGID DESIGNATOR

Rigid designation is a notion invented by Kripke and a rigid designator is a term that refers to the same object in all possible worlds in which it refers at all. Proper names are rigid designators. Definite descriptions are not, *qua* descriptions, rigid. “The current prime minister of the U K” could have designated someone other than David Cameron, and does so in other possible

worlds. On the other hand, it can be used rigidly — “the current prime minister of the U K—that is, David Cameron.” When used rigidly, you can say “the current prime minister might not have been the prime minister”—which would have been true if the Conservatives had lost the last election and Cameron had not become prime minister. On the other hand, “David Cameron might not have been David Cameron” makes no clear sense.

108 SATISFY

An object *satisfies* a predicate when the predicate is true of it. The term is employed in Tarski’s theory of truth to name the relation that holds between predicates and objects that fall under them, and the notion of truth in the language to which the predicates belong is exhibited by generating the truth conditions for the language via these initial assignments and the grammar of the language.

109 SATURATED

According to Frege, the difference between an object and a concept is that the former is saturated and the latter is not. The form of a concept expression is “. . . is F,” and it is unsaturated—incomplete—because of the empty argument place. An object expression is the name or referring expression that occupies such an argument place, and its form is “a” and is not an incomplete expression, hence it is *saturated*. Although this is true of the expressions in question, it does not follow that objects can exist or be complete without properties, any more than properties can exist without objects, so it is not clear how deep metaphysically the distinction in logical form goes. Frege thought it was a deep fact because he regarded concepts as analogous to mathematical functions, which are essentially operators on something else, ultimately numbers, which are self-standing entities.

110 SCOPE

The scope of an operator is the range of its coverage, which might be wide or narrow. The following are examples. In the case of the quantifier shift fallacy (see above), if one says “there is something which is the cause of everything,” then the existential quantifier has the wide scope because it comes first and covers the whole expression; whereas in “everything has some cause or other” the universal quantifier has the wider scope. In the case of rigid designation (see above), “the prime minister might not have been the prime

minister” makes good sense when read as “there is something that is the prime minister and it is possible that that thing not be the prime minister.” Here the existential quantifier has wider scope than the possibility operator. The nonsensical version is “it is possible that there is something which is the prime minister and is not the prime minister,” where the possibility operator has the wider scope. In logical formulae the scope is usually expressed by bracketing.

111 SEMANTIC THEORY OF TRUTH

This is a theory associated with Tarski. The objective is to define “truth” for a given language (natural or formal) by a system that generates the so-called *T-sentences* for the language. Such a theory for English would generate equivalences of the form

“snow is white” is true if and only if snow is white

for all well-formed indicative sentences of English. As it gives the truth conditions for sentences it can be regarded as a deflationary version of the correspondence theory of truth: deflationary because no further account of the correspondence relation than the equivalence is deemed necessary.

112 SEMANTICS AND SYNTAX

The semantics of linguistic elements concerns how they relate to the world and the syntax concerns their structural relation to each other. So the semantics of individual words is very roughly equivalent to the thing or things to which they refer or which they cover—their extension—and the syntax of sentences is the grammar that determines which combination of words is “well-formed” and makes structural sense. Both categories apply to both natural and formal languages.

113 SENSE AND REFERENCE

These terms are used to translate Frege’s *Sinn und Bedeutung*. The latter is what an expression denotes, the former is the meaning or descriptive content of a term. When an expression has both features it is often said that the sense is the *mode of presentation* of the thing denoted. Thus “the morning star” refers to Venus, but it presents Venus as that heavenly body that appears in the morning.

114 SENSE-DATA

This is a term that was introduced in the early twentieth century to denote the immediate contents of sense-experience, that is, the sensible qualities of which we are aware. G. E. Moore initially considered whether the visual sense-data might be the colored surfaces of objects, but concluded, on the grounds of the argument from illusion, that they could not be. Given this conclusion, sense-data are roughly equivalent to the *ideas* of Locke and Berkeley, or Hume's *impressions*. The fundamental rationale behind the theory is that, if one clearly seems to see something, for example, red, then one is aware of something red, even if there is no relevant external red object. One is, therefore, aware of a red sense-datum. The crux is the claim that red is really present, not just as an intentional object. The "sense-datum theory of perception" was very popular until after 1945, when there was an onslaught from "common sense" and "ordinary language" philosophy, and various versions of the intentional theory, which led to the almost complete abandonment of the theory. More recently, it has become clear that the argument from illusion and the causal-hallucinatory arguments are not easily dismissed.

115 SENTENCES, PROPOSITIONS, STATEMENTS

A sentence is a grammatically well-formed expression in a language. A proposition is what is said or asserted in an indicative sentence. The term "statement" is sometimes used as equivalent to "proposition" and sometimes used to designate the act of asserting a proposition. The sentence/proposition distinction is generally regarded as necessary because the same sentence can be used to assert different propositions on different occasions. This is true, for example, of sentences with indexicals or demonstratives. "It is raining here," for example, asserts different propositions of different places and times. Quine wished to eliminate propositions by employing "universals sentences," that is, sentences from which all indexical elements have been eliminated. It is not generally thought that this can always be done. The distinction between sentences and propositions is also needed because not all sentences are indicatives; some are questions and some imperatives, for example. These do not assert any proposition at all, and there is no term parallel to "proposition" showing what they do—they are just questions and imperatives.

116 SET

A collection of entities, which are the members or elements of the set. The collection of these entities—the set—is viewed as an object in its own right, which

exists in addition to the elements. A set is individuated by its members, so it is an *extensional* notion: sets with the same members are identical. The “union” of two sets consists of the members of both. The “intersection” of two sets is the entities that are members of both. Whether sets understood in this manner really exist, or are merely a fiction devised by mathematicians, is an issue that divides metaphysicians.

117 SINN

See *Sense and Reference*.

118 SORTAL

The term “sortal” was invented by Locke to characterize any *type of thing*, where thing means an object of the sort picked out by a common noun, not an abstraction or a property. The term was taken up by P. F. Strawson, but used most systematically by David Wiggins. Wiggins uses it so as to coincide as nearly as possible with what Aristotle would describe as a substance. A true sortal applies to an object through the whole of its existence and provides its identity conditions. No object can fall under two different sortals, for no object can have two different sets of identity conditions, upon pain of contradiction.

119 SPEECH ACT THEORY

A theory of the nature of language derived from the work of J. L. Austin and developed particularly by John Searle. The inspiration comes from Austin’s volume *How To Do Things With Words* and the central idea is that a language is not a formal or abstract system, nor a machine for generating true statements but a tool we use to achieve many different things. In particular, Austin distinguished *locutions*, which are straightforward statements of fact, *illocutions* by which we perform a nondescriptive act, and *perlocutions* by which we cause something to happen. The classic illustration is the use of the sentence “there is a bull in the field” uttered by A to B while B is in the field in question. A *describes* a certain state of affairs, he *warns* B and he *causes* B to leave the field by the safest route available. Searle famously applied these tools to an analysis of promising, which he thought helped overcome the traditional is-ought dichotomy.

120 STATEMENT

See *Sentences, Propositions, Statements*.

121 SUPERVENIENCE

One collection of entities (states of affairs or properties) or predicates *supervenenes* on another collection if and only if the variation in the former cannot occur without variation in the latter. So, for example, two pictures cannot be different in their aesthetic qualities unless there is some difference in their physical properties: they cannot be physically just alike except that one is more beautiful than the other. Similarly, two actions (in exactly similar contexts) cannot differ in moral properties unless they differ in some physical way. Since Davidson's 1970 book, the term has mainly been applied to the philosophy of mind, by physicalists who do not want to commit themselves to reduction of the mental to the physical. They say that the mental supervenes on the physical, so two mental situations cannot vary unless there is also some physical difference. There could not, for example, be a philosophers' zombie—a creature exactly like a conscious human being physically, yet without consciousness. The status of the necessity in question is not always clear.

122 TIME: A-SERIES AND B-SERIES

The idealist philosopher J. M. E. McTaggart distinguished between two kinds of temporal relations, the *A-series* that consists in the properties past-present-future, and the *B-series*, which consists in the relations of being earlier than, simultaneous with, and later than. The salient difference is that the former properties change and the latter do not. So a given event is first in the future, then, as it occurs, is in the present, then when it has occurred, is in the past. Events, however, never change their B-series relations to each other: if, for example, event x is earlier than event y, then it is so whether they are now in the past, present, or future. McTaggart argued that (i) time requires the constantly changing A-series, properties, (ii) the A-series is inherently contradictory, so he drew the conclusion that (iii) time is unreal. Analytic philosophers have not followed this path, but they regard the distinction between the series as important and four-dimensionalists want to deny the reality of the A-series, dubbing it "the myth of passage." See *Four-dimensionalism, Growing Block*, and *Presentism*.

123 TRUTH CONDITION

The truth condition or conditions of a proposition is/are those things (situations, states of affairs) the holding of which make it to be true. The "make" in

question is not causal, but constitutive—its being true is just for those conditions to obtain.

124 TRUTH-FUNCTIONAL

A logical connective is truth functional if and only if the truth value of the complex proposition it governs is a mechanical function of the truth values of the propositions it connects.

125 TRUTH-VALUE GAPS

The truth value of a proposition in classical logic is either “true” or “false,” but in some nonclassical logics some propositions can fail to be either true or false and have an indeterminate truth value. For example, a proposition using a vague predicate applied in a case where the vagueness comes into play, may be neither true nor false. So “Fred is bald” said of someone who is neither bald nor hirsute is, according to such a logic, neither true nor false, so there is a truth value gap.

126 TYPES VERSUS TOKENS

How many words are there in the sentence “all men are men”? It depends on whether one means word-types or word-tokens. There are four word tokens but only three word types, as the word “men” occurs twice and so is one type with two tokens.

127 UNIVERSALIZABILITY

A principle is universalizable if it applies to all cases falling under the same description; mere numerical difference cannot *make* a difference. Moral rules are generally held to be universalizable, because some action cannot be right if I do it but wrong if you do, unless there is some general characteristic that differs in the two cases, so making our actions fall under slightly different principles. Those who do not believe that moral rules are universalizable are *particularists*.

128 UNIVERSALS

See *Nominalism*.

129 USE-MENTION DISTINCTION

In the sentence the “cat is on the mat” the word “cat” is being *used*, but in the sentence “the word ‘cat’ has three letters” it is being *mentioned*. Words are mentioned when they are not employed in their normal work of “being about” the world, but are being quoted so that the word itself is what is being referred to. It has been claimed that certain philosophical mistakes have been made from the failure to make this distinction.

130 VAGUENESS

Some concepts are vague, so that there seems to be no clear division between the cases where they apply and where they do not. Standard examples include the concept *bald* and the concept *mountain*. Some people are clearly bald, some clearly not, but there are many cases of people who neither clearly are nor clearly not. Similarly, for whether something counts as a mountain. One consequence of this has been taken by some philosophers to be that classical two-valued logic is inadequate, for as well as the values *true* and *false*, we also need a third value representing neither true nor false.

Vagueness can also lead to the *sorites paradox* otherwise known as the *paradox of the heap*. (“Sorites” is the Greek word for “heap.”) A hundred grains of sand piled together constitutes a heap, but the removal of one grain from a heap is never enough to render it not a heap. Applied recursively, this latter principle leads to the conclusion that one grain and no grains still constitute a heap, which is absurd. Similar moves can be made for *bald* (removing one hair never renders a hairy person bald . . .) and *mountain* (if something at n feet is a mountain, so is something at n minus 1 feet . . .) and for many vague predicates.

Notice that this kind of vagueness is not the same as indeterminacy: one cannot work the sorites on, for example, quantum indeterminacy. Many philosophers claim that sorites vagueness is a conceptual phenomenon, not a feature of reality (unlike quantum indeterminacy). Others claim that vague predicates really do have a cut off point, we are just unable to discern it.

131 WIDE VERSUS NARROW CONTENT

A mental state has narrow content if and only if it is, or is constituted or individuated by, a situation entirely internal to the subject who has it. A mental state has wide content if the mental state is constituted or individuated by a situation that involves something external to the subject. According to an externalist about natural kinds, someone can have our concept WATER only

if they live in a world containing H_2O and their concept is causally connected to the H_2O . If they had lived in a world in which there was watery stuff around but it was differently constituted—for example, by XYZ—they would have a different concept, even though they might be internally just as we are. An internalist, on the other hand, believes that a brain in a vat could have the concept WATER provided it had watery stuff type experiences, just like our experiences: the internal experiential state is enough to ground possession of the concept. Some philosophers believe that we can have concepts with both wide and narrow content. So our fully-fledged natural kind concept WATER requires the presence of H_2O , and so is wide, but someone in the XYZ world, or even in a vat with the right experiential states could share with us the more modest concept WATERY STUFF. David Chalmers claims that our concept WATER itself has both these dimensions: the narrow content is the primary intension, the wide content the secondary intension.

Resources

(a) The online resource *philpapers* is (in its own words) “a comprehensive directory of online philosophical articles and books by academic philosophers” and can be found here:

<http://philpapers.org/>

The site allows one to monitor current research by browsing new issues of over 350 journals, and explore the previous literature via the extensive categorization system and advanced search engine, which quickly allows one to find out all the publications of a given philosopher and (very often) to gain access to their publications online. The resources to which the site provides access are vast and always expanding; at the time of writing there are some 82,000 entries in metaphysics and epistemology, 125,000 in ethics, 80,000 in science, logic, and mathematics, and 90,000 in the history of Western philosophy.

Two related sites are philjobs and philevents:

<http://philevents.org/> and <http://philjobs.org/>

Their purpose is self-explanatory.

(b) The online *Stanford Encyclopedia of Philosophy* is a widely used and authoritative resource:

<http://plato.stanford.edu/>

There are useful entries on hundreds of relevant topics to be found here, including all the major philosophers (Russell, Wittgenstein, Moore, Bradley, etc.), names, theory of descriptions, logical constructions, possible worlds, nonexistent objects, axiomatic theories of truth (along with coherence, deflationary, and pluralist theories), behaviorism, substance, consciousness, functionalism, and so on.

(c) The *Internet Encyclopedia of Philosophy* has useful entries on many topics, including the history of analytic philosophy:

www.iep.utm.edu/category/history/analytic/

- (d) The “hist-analytic” Website is also worth noting:
www.hist-analytic.com/

It brings together resources useful to historians of analytic philosophy, including writings from Ayer, Aune, Reichenbach, Mach, Russell, Price, Passmore, Hintikka, Feigl, and others.

- (e) The Website for *The Journal for the History of Analytical Philosophy* can be found here:

<http://jhaponline.org/journals/jhap/index>

The journal “aims to promote research in and provide a forum for discussion of the history of analytical philosophy. ‘History’ and ‘analytic’ are understood broadly. JHAP takes the history of analytic philosophy to be part of analytic philosophy. Accordingly it publishes research that interacts with ongoing concerns of analytic philosophy and with the history of twentieth century philosophical traditions.”

- (f) The site of the *Cambridge Wittgenstein Archive* can be found here:
www.wittgen-cam.ac.uk/

It is well worth a visit; among other things, includes a useful and well-illustrated online biography of Wittgenstein.

- (g) The *Bertrand Russell Research Centre* site can be found here:
www.humanities.mcmaster.ca/~russell/

This is a useful source of Russell-related news; there is also a biography, gallery, and access to some of Russell’s writings.

- (h) Another site worth visiting is the *MacTutor History of Mathematics* archive:
<http://www-history.mcs.st-andrews.ac.uk/>

This is an online resource, with useful articles on many topics in maths and logic, and biographies, as well as links to further resources.

- (i) There are a good many blogs devoted to analytic philosophy and philosophers, and they are a good way of seeing analytic philosophers in action, finding out what the hot issues currently are—and keeping up with the gossip. The blog scene changes quickly.
<http://consc.net/weblogs.html>

Bill Valicella’s

http://maverickphilosopher.typepad.com/maverick_philosopher/

Resources

(j) Philosophy e-mail lists are a useful way of keeping track of happenings, discussions, etc. Two of the most widely used are:

PHILOS-L—to subscribe send a mail to LISTSERVE@LIV.AC.UK with the command in the email body `SUBSCRIBE PHILOS-L`

PHILOSOP (<http://philosophy.louisiana.edu/philosop.html>)

For a more complete list see:

<http://homepages.ed.ac.uk/ejua35/mailgen.htm>

Annotated Bibliography

1 Part A—Analytic Philosophy and Its History

1.1 The Founding Fathers

1.1.1 Gottlob Frege

Beaney, M., ed., 1997. *The Frege Reader*. Oxford: Blackwell.

This volume contains almost all of Frege's philosophically essential writings, including "On concept and object," "On sense and reference," and "Thought."

Kenny, A., 2000. *Frege: An Introduction to the Founder of Modern Analytic Philosophy*. Oxford: Blackwell.

Kenny is always one of the most lucid historians of philosophy and this is particularly important when trying to get to grips with Frege.

Dummett, M., 1973. *Frege: Philosophy of Language*. London: Duckworth (Cambridge, MA: Harvard University Press, 1991).

This has been a classic ever since it was published, but is definitely not the place to begin. Dummett is probably the most influential expositor of Frege, but he writes at great length with few signposts to help the reader.

1.1.2 Bertrand Russell

Marsh, R. C., ed., 1956. *Logic and Knowledge: Essays 1901–1950*. London and New York: George Allen and Unwin.

There are various collections of Russell's essays published at various times in his life, but this covers the widest timeframe. It includes "On denoting" and the essays on logical atomism. It is more true of Russell than of Frege that one should read his books as well as his essays. The simplest is *Problems of Philosophy*, in many editions. Many of his other works can be found in the main bibliography.

Sainsbury, M., 1979. *Russell*. London: Routledge.

This book in the “Arguments of the Philosophers” series contains first-rate discussions of Russell’s philosophy. Although the issues are often not easy, Sainsbury is a very clear writer, and this is a very useful book for anyone trying to come to grips with Russell’s development.

Candlish, S., 2007. *The Russell/Bradley Dispute and Its Significance for Twentieth-century Philosophy*. Basingstoke and New York: Palgrave Macmillan.

A valuable guide to issues that divided Russell and Bradley—such as relations—and Russell’s changing views on propositional unity. Candlish argues (persuasively) that the stereotypical view among analytic philosophers of the relationship between Russell and Bradley—particularly with regard to the success of the former’s criticism of the latter—is erroneous in several respects.

Monk, R., 1996. *Bertrand Russell: The Spirit of Solitude*. London: Random House.

An excellent biography of Russell, covering the part of his long career that is most relevant to his contribution to analytic philosophy. Russell’s own *Autobiography* (Routledge, 2009, but available in many previous editions) is also very readable and extremely interesting.

1.1.3 G. E. Moore

Moore, G. E., 1903. *Principia Ethica*. Cambridge: CUP.

This book was regarded for much of the twentieth century as the beginning of modern moral philosophy. Although now regarded as somewhat less central, it is not possible to understand analytical ethics, nor Moore’s importance, without it.

Moore, G. E., 1959. *Philosophical Papers*. London: George Allen and Unwin.

This collection, put together by Moore just before his death, and published just after it, contains most of Moore’s major contributions to epistemology and the philosophy of language.

Baldwin, T., 1990. *G. E. Moore*. London: Routledge.

There are not many monographs devoted to Moore alone and this volume in the “Arguments of the Philosophers” series is a major contribution.

1.1.4 Ludwig Wittgenstein

Wittgenstein, L., 1922/1961. *Tractatus Logico-Philosophicus*. London: Routledge and Kegan Paul.

This is logical atomism without the (explicit, at least) empiricism of Russell. Despite its obscurity, its influence on logical positivism gives it a fundamental importance.

Wittgenstein, L., 1967. *Philosophical Investigations*. Translated by G. E. M. Anscombe. Oxford: Blackwells.

The influence of this work on “ordinary language philosophy” and on the anti-empiricism and anti-Cartesianism that has marked much modern philosophy can hardly be overstated, although this is less true in the United States than in Britain. Wittgenstein is the most controversial of the founders of analytic philosophy. Many regard him as the greatest philosophical genius of the twentieth century, but for some he is little short of a charlatan.

Ayer, A. J., 1985. *Wittgenstein*: London: Weidenfeld and Nicholson.

Kenny, A. 2006. *Wittgenstein* (revised edition), Oxford: Wiley-Blackwell).

Two accessible, and reasonably brief, guides to both early and late Wittgenstein. Since Ayer and Kenny have very different views as to which of Wittgenstein’s views stand up to scrutiny, those new to the literature can benefit from reading both books. For an entertaining investigation into the originality of Wittgenstein’s early work, see

Goldstein, L., 2002. “How Original a Work Is the *Tractatus Logico-Philosophicus*?” *Philosophy*, 77, pp. 421–46.

For absolutely committed believers in Wittgenstein’s later philosophy, see the four volumes given below.

Hacker, P. and Baker, G., 1980, 1985, 1990, 1996. *Analytic Commentary on the Philosophical Investigations*: Oxford: Basil Blackwell.

The most detailed commentary on the later Wittgenstein’s work, by two leading Wittgensteinians, extending over four substantial volumes—the latter two by Hacker alone.

Kripke, S., 1982. *Wittgenstein on Rules and Private Language*. Cambridge, MA: Harvard University Press.

This short book by one of the world’s leading living analytic philosophers electrified the philosophical community—or that part of it interested in Wittgenstein—when it appeared in 1982. Kripke’s reinterpretation of the central doctrines of the *Investigations* is controversial, but well worth reading.

1.2 Later Movements

1.2.1 Logical Positivism

Ayer, A. J., 1936. *Language, Truth and Logic*. London: Gollancz, 1936 (2nd edn with new introduction, 1945); also Ayer, A. J., ed., 1959. *Logical Positivism*. London: George Allen and Unwin.

Given their ideology and their background in logic, mathematics, and physics, many of the leading positivists—Carnap, for example—tend to move rather

swiftly from general principles to the purported technical implementation of those principles. Ayer, who had no formal background and thought in the careful analytical way of the British empiricists, spent more time in trying to articulate and develop the *ideas* behind positivism and thus created a much more accessible and discussable philosophical exposition of positivism than did the others. This is the strength of *Language, Truth and Logic*. Similarly, the collection Ayer edited concentrates on the ideas of the group, not their technicalities.

Richardson, A. and Uebel, T., eds, 2007. *The Cambridge Companion to Logical Empiricism*. Cambridge: CUP.

In recent years “Logical Empiricism” has been the favored label among many experts for the movement initiated by the Vienna Circle, which also goes by the name “Logical Positivism.” This is a useful collection of recent essays, focusing on the historical context and formative stages of logical empiricism, its main achievements (and failures) in the philosophy of science, and its influence on philosophy past, present, and future.

Friedman, M., 1999. *Reconsidering Logical Positivism*. Cambridge: CUP.

A collection of papers by one of the leading “reinterpreters” of logical positivism. Friedman argues that the positivists in general, and Carnap in particular, can only properly be understood—and appreciated—if they are considered in their historical framework, specifically, the various neo-Kantian traditions that dominated the German philosophical world at the turn of the twentieth century.

1.2.2 Ordinary Language Philosophy

Austin, J. L., 1962a. *How to Do Things with Words: The William James Lectures, 1955*. Oxford: Clarendon.

—, 1962b. *Philosophical Papers*. Oxford: OUP.

Together these works show “ordinary language philosophy,” Oxford version, from the pen of its most influential exponent. The former of the two was Austin’s most systematic and enduring contribution to philosophy, as it gave rise to “speech act theory.”

Searle, J., 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: CUP.

This was the first major attempt to develop Austin’s approach to language in a systematic way. It also established John Searle’s importance as a philosopher.

Strawson, P. F., 1971. *Logico-Linguistic Papers*. London: Methuen.

Strawson was both one of the most original and influential philosophers who developed within the “Oxford” tradition of the 1950s, and the major figure in a move back to a form of metaphysics. These papers, however, include his famous attack on Russell’s theory of descriptions and his natural, language-based resistance to the domination of formality in the philosophy of language.

Grice, P., 1989. *Studies in the Way of Words*. Cambridge MA: Harvard University Press.

Grice was more minute in his manner of philosophizing than Searle, and was responsible for developing a systematic theory of meaning within a “speech act” framework.

1.3 Analytic Philosophy as a General Topic

Urmson, J. O., 1956. *Philosophical Analysis: Its Development between the Wars*. Oxford: OUP.

Urmson’s typically brief work is, characteristically, a paradigm of good sense and lucidity. It is one of the earliest works reflecting on the tradition and still repays reading.

Beaney, M., ed., 2007. *The Analytic Turn: Analysis in Early Analytic Philosophy*. London: Routledge.

For anyone seeking to know more about what the “analysis” in analytic philosophy amounts to.

Hylton, P., 1990. *Russell, Idealism, and the Emergence of Analytic Philosophy*. Oxford: Clarendon.

Perhaps the best critical survey of early analytic philosophy, and Russell’s role in it. Hylton examines the arguments, and motivations behind them, in considerable detail.

Coffa, A. J., 1991. *The Semantic Tradition from Kant to Carnap*. Cambridge: CUP.

Coffa argues that both logicism and positivism should be viewed as parts of a “semantic tradition” stretching back to Bolzano and other nineteenth-century responses to—and rejections of—Kant’s views on the a priori.

Martinich, A. P. and Sosa, D., eds, 2001. *A Companion to Analytic Philosophy*. Oxford: Blackwell.

Essays on 41 philosophers, including Frege, Moore, Russell, Wittgenstein, Broad, Carnap, Ryle, Hempel, Goodman, Hart, Stevenson, Foot, Strawson, Anscombe, Chisholm, Quine, Rawls, Kuhn, Dummett, Armstrong, Chomsky, Searle, Fodor, Putnam, Kripke, and Lewis.

Hacker, P. M. S., 1996. *Wittgenstein’s Place in Twentieth-century Philosophy*. Oxford: Blackwell.

Although Hacker’s primary focus is on Wittgenstein’s influence, he provides useful expositions of the doctrines of both Wittgenstein and many other leading analytic philosophers.

Schwartz, S. P., 2010. *A Brief History of Analytic Philosophy: From Russell to Rawls*. Oxford: Wiley-Blackwell.

A very useful and accessible guide to all the main early figures and movements; there are also chapters on more recent work in mind, language, and metaphysics, and an epilogue, including “Analytic Philosophy since 1980” and “What Is the Future of Analytic Philosophy?”

Soames, S., 2003. *Philosophical Analysis in the Twentieth Century, Volume 1: The Dawn of Analysis, Volume II: The Age of Meaning*. Princeton, NJ, Princeton University Press.

Soames covers a lot of ground, and in a distinctive way: his main concern is with assessing the quality of the arguments he deals with, rather than in tracing their historical connections and motivations, or providing comprehensive introductions to the views of the philosophers he discusses. These two volumes thus combine an overview of many of the most influential doctrines in analytic philosophy, and also demonstrate what it is to *do* analytic philosophy.

Dummett, M., 1993. *The Origins of Analytic Philosophy*. London: Duckworth.

Dummett takes the strongest and narrowest conception of analytic philosophy of any recent major figure. Proper philosophy is about the study of thought by means of the study of language, and Frege invented it.

Williamson, T., 2007. *The Philosophy of Philosophy*. Oxford: Blackwell.

Williamson is probably the most influential philosopher in the analytic tradition in Britain in 2013, and in this book he tries to work out what philosophy is. He distinguishes himself from Dummett’s rather narrow conception but castigates the profession for their generally lax standards!

Glock, H.-J., 2008. *What is Analytic Philosophy?* Cambridge: CUP.

This splendidly readable work is historically informative, persistent in its attempts to answer the question that is its title, and often very amusing.

2 Part B—By Subject and Chapter

2.1 The Philosophy of Mathematics

Philosophy of Mathematics: Selected Readings. 2nd edn, 1983, introduced and edited by Paul Benacerraf and Hilary Putnam. Cambridge: CUP.

This book contains an excellent historical selection, including Carnap, Frege, von Neumann, Brouwer, Russell, Hilbert, Quine, Goodman, Hempel, Parsons, Gödel, Dummett, and important contributions from the editors themselves.

Shapiro, S., 2000. *Thinking about Mathematics: The Philosophy of Mathematics*. Oxford: OUP.

A useful and accessible introduction, with sections devoted to an historical overview going back to Plato and the Greeks, the major positions held in the twentieth century, and a final section devoted to contemporary developments.

Colyvan, M., 2012. *An Introduction to the Philosophy of Mathematics*. Cambridge: CUP.

Another useful overview, with a particular focus on the limits and (to some) mysterious applicability of mathematics to the real world; the “epilogue” of “desert island theorems” provides a brief introduction to many of the theorems of maths and logic that philosophers are most fond of, plus several that the author believes are underappreciated.

Shapiro, S., ed., 2005. *The Oxford Handbook of Mathematics and Logic*. Oxford: OUP.

After a section devoted to relevant historical background (with chapters on empiricism, Kant, and Wittgenstein)—itself preceded by a general overview of the field(s) by the editor—there are sections devoted to logicism, formalism, intuitionism, structuralism and nominalism, and logical consequence.

Papineau, D., 2012. *Philosophical Devices: Proofs, Probabilities, Possibilities and Sets*. Oxford: OUP.

Steinhart, E., 2009. *More Precisely: The Math You Need to Do Philosophy*. Peterborough, Ontario: Broadview.

A background knowledge of set theory, logic, formal and possible worlds semantics, and probability theory is presupposed in many areas of contemporary analytic philosophy: in the philosophy of mathematics certainly, but also the philosophies of language, mind, and science—even ethics and political philosophy. These books have similar aims: to introduce newcomers to the technical terms and concepts that they can expect to encounter. Both are very useful indeed.

2.2 The Philosophy of Language

The works cited in Part (A) by Austin, Searle, Strawson, and Grice are seminal. Miller, A., 2007. *Philosophy of Language*. 2nd edn. London: Routledge.

This useful textbook covers many of the central topics—for example, Frege and Russell, logical positivism, Quine on analyticity, Kripke’s Wittgenstein, Grice, Davidson, and Tarski—in an accessible manner.

Ludlow, P., ed., 1997. *Readings in the Philosophy of Language*. Cambridge MA: MIT Press.

Martinich, A. P., ed., 2008. *The Philosophy of Language*. 5th edn. Oxford: OUP.

These are both excellent and comprehensive anthologies, containing original articles by leading contributors to the field.

- Kripke, S., 1982. *Wittgenstein on Rules and Private Language*. New York: Harvard University Press.
—, 1980. *Naming and Necessity*. Oxford: Blackwell.

Two of the most influential monographs to have been published in recent decades, both eminently readable.

2.3 The Philosophy of Science

- Ladyman, J., 2002. *Understanding Philosophy of Science*. London: Routledge.
Bird, A., 1998. *Philosophy of Science*. London: Routledge.
Hacking, I., 1983. *Representing and Intervening*. Cambridge: CUP.

These number among the many excellent introductions to the philosophy of science.

- Lange, M., ed., 2007. *Philosophy of Science*. Oxford: Blackwell.
Bird, A. and Ladyman, J., eds, 2012. *Arguing about Science*. London: Routledge.

These are both up-to-date anthologies that contain many of the classic articles by leading philosophers of science, together with more recent contributions.

- Kuhn, T., 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

The book that introduced “paradigm” into the philosophy of science, Kuhn’s book is probably the most-discussed single-work in the philosophy of science in recent decades.

- Popper, K., 1959. *Conjectures and Refutations*. London: Routledge.

Popper was among the most influential philosophers of science for many years; this book is a good introduction to his thinking in a number of areas.

2.4 The Philosophy of Physics

- Sklar, L., 2002. *Philosophy of Physics*. Oxford: OUP.

The philosophy of physics literature can be dauntingly technical for the nonspecialist, but Sklar here provides an accessible and nontechnical introduction to the main issues, focusing on three areas: (a) space, time, and motion, (b) the introduction of probability into physics via statistic mechanics, and (c) the problems and puzzles to which quantum theory has given rise.

- Dainton, B., 2010. *Time and Space*. 2nd edn. Durham: Acumen.

Focuses on the main metaphysical issues to which time and space give rise—Is space a thing? Does time pass?—and examines the way in which developments in physics and mathematics have impacted on these.

Price, H., 1996. *Time's Arrow and Archimedes' Point*. Oxford: OUP.

A more advanced, but highly interesting and provocative, treatment of temporal asymmetries and quantum theory.

Batterman, R., ed., 2013. *Oxford Handbook of Philosophy of Physics*. Oxford: OUP.

This collection contains chapters on contemporary debates on causation, classical mechanics, symmetry, phase transitions, unification, field theories, spacetime structure, interpretations of quantum theory, and cosmology.

Albert, D., 1992. *Quantum Mechanics and Experience*. New York: Harvard University Press.

An accessible, but deep, introduction to the central issues to which quantum mechanics gives rise, and some of the leading solutions to the "measurement problem."

Maudlin, T., 2007. *The Metaphysics within Physics*. Oxford: OUP.

Here Maudlin defends the thesis that metaphysics "insofar as it is concerned with the natural world, can do no better than reflect on physics," that is, in developing our metaphysics we should pay attention primarily to the account of the fundamental features of the world that are implicit in our best physical theories.

Butterfield, J. and Earman, J., eds, 2007. *Philosophy of Physics*. Amsterdam: Elsevier.

A comprehensive—but very advanced—survey of the central issues by leading contemporary theorists, containing a useful (and accessible) introduction.

2.5 Causation

Lewis, D., 2000. "Causation as influence." *The Journal of Philosophy*, 97, pp. 182–97. Reprinted in Collins, J., Hall, N., and Paul, L. A., eds, 2004. *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

This is David Lewis' final attempt to construct a viable counterfactual analysis of causation. It defines causation not in terms of "whether-whether" dependence (so that whether or not the effect occurs counterfactually depends on whether the cause occurs), but rather in terms of "influence" or the extent to which "alterations" of the effect counterfactually depend on "alterations" of the cause. Thus *c influences e* to the extent that whether, how, and when *e* occurs counterfactually depends on whether, how, and when *c* occurs.

Beebe, H., Hitchcock, C., and Menzies, P., eds, 2009. *The Oxford Handbook of Causation*. Oxford: OUP.

This is a comprehensive collection aiming to summarize both the history of the causation debate and the current state of the many contemporary debates, including the role of causation in philosophical theories more widely and in other disciplines.

Hall, N., 2004. "Two Concepts of Causation," in Collins, Hall, and Paul 2004, pp. 225–76.

This paper puts forward the influential thesis that there are two concepts of causation rather than one: a "dependence" concept and a "production" concept. A large part of its interest rests on its diagnosis of, and suggested solution to, many of the issues surrounding the analysis of causation that have proved intractable in recent years.

Anscombe, G. E. M., 1971. *Causality and Determination: An Inaugural Lecture*. Cambridge, MA: CUP. Reprinted in Sosa, E. and Tooley, M., eds, 1993. *Causation*. Oxford: OUP.

Anscombe's classic lecture aims to sound the death knell on Hume's influential claim that causation is to be understood in terms of regularity.

2.6 Metaphysics

Loux, M. J., 2006. *Metaphysics: A Contemporary Introduction*. 3rd edn. New York: Routledge.

Lowe, E. J., 2002. *A Survey of Metaphysics*. Oxford: OUP.

There are plenty of textbooks that offer good introductions to the main contemporary debates in analytic metaphysics, but these are two of the better ones.

van Inwagen, P. and Zimmerman D., eds, 2008. *Metaphysics: The Big Questions*. 2nd edn. Oxford: Wiley-Blackwell.

There is also no shortage of impressively weighty anthologies containing selections of the most influential original articles in analytic metaphysics, but this covers a wider range of interesting territory than most. There are 65 articles in total, distributed over four sections (a) What Are the Most General Features of the World? (b) What Is Our Place in the World? (c) Are There Many Worlds? And (d) Why Is There a World?

Lewis, D., 1986. *The Plurality of Worlds*. Oxford: Basil Blackwell.

van Inwagen, P., 1990. *Material Beings*. Ithaca, NY: Cornell University Press.

These are two of the most influential monographs in metaphysics of recent decades.

Chalmers, D., Manley, D., and Wasserman, R., eds, 2009. *Metametaphysics: New Essays on the Foundations of Ontology*. Oxford: OUP.

The recent resurgence of metaphysics in analytic philosophy has led to a resurgence of interest in the nature of metaphysics itself. For anyone interested in these debates, this anthology is a good place to start.

2.7 The Philosophy of Mind

Broad, C. D., 1925. *The Mind and its Place in Nature*. London: Kegan Paul, Trench, Trubner.

After some decades of neglect this masterpiece by one of the greatest of the early twentieth-century Cambridge philosophers has returned to the status of a classic. Broad's careful sorting out of all options—he distinguishes between 17 different theories of mind-body relations—and his lack of partisan arrogance marks him off from many of his contemporaries and successors.

Ryle, G., 1949. *The Concept of Mind*. London: Hutchinson.

Despite its tendentiousness and the weakness of most of its central arguments, Ryle's classic is essential reading because of the pervasiveness of its influence. Ryle managed, for some decades, to convince the philosophical community that he had exorcised the inner, private realm of the mental and replaced it, for the most part, with an informal behaviorism. This provided the background for Armstrong's form of central state materialism and Dennett's instrumentalist theory of the mind.

Armstrong, D. M., 1968. *A Materialist Theory of the Mind*. London: Routledge and Kegan Paul.

Armstrong's central state materialism sought to make the mind less elusive or abstract than behavioral dispositions, while preserving what is taken to be the insight of behaviorism, namely that mentality consists essentially in its contribution to behavior. Whether this central state "causal theory of the mind" actually makes the improvements on behaviorism that it thinks it does, is controversial. But Armstrong's theory has much in common with functionalism and the book is typical of his work in the lucid and fair-minded way his position is developed. Despite appearing to have been displaced by later works by other authors, Armstrong's book remains a classic statement of how physicalism would need to work.

Chalmers, D., 1996. *The Conscious Mind*. Oxford: OUP.

Chalmers's book became an immediate classic, for at least three reasons. First, it contains an architectonic account of the options not unlike Broad's (though fewer in number than Broad's 17!). Second, it develops a very powerful case against reductive physicalist accounts of consciousness, drawing on recent

innovations in modal semantics. Third, Chalmers explores some novel (and not so novel) accounts of the relationship between consciousness and physical reality, including a new version of property dualism, and a form of neutral monism-cum-pan-psychism, bringing to the center of discussion a view that had previously been regarded as too outrageous to be taken seriously. It is a book that no one can ignore.

There are a number of comprehensive anthologies of classic readings in the philosophy of mind, one of the best and most up-to-date is:

Lycan, W. G. and Prinz, J. J., eds, 2008. *Mind and Cognition: An Anthology*. Oxford: Wiley-Blackwell.

2.8 Personal Identity

Parfit, D., 1984. *Reasons and Persons*. Oxford: OUP.

It is part three of this book that has had so much influence on the philosophy of personal identity. There Parfit argues for the view that when it comes to personal survival what really matters, in the sense of what it is rational to care about, is not identity, but its usual concomitant, namely psychological continuity. The discussion crucially depends on what has come to be known as the "argument from below."

Olson, E., 1997. *The Human Animal: Personal Identity without Psychology*. Oxford: OUP.

Offers a vigorous defense of the view that we are animals, that is, organisms of a particular sort. Olson's master argument against his opponents is that they postulate too many thinkers.

Martin, R. and Barresi, J., eds, 2003. *Personal Identity*. Oxford: Blackwell.

Still the best anthology on personal identity, with important papers by Bernard Williams, Robert Nozick, David Lewis, and others.

Johnston, M., 2010. *Surviving Death*. Princeton, NJ: Princeton University Press.

Provides an entirely novel account of personal identity.

2.9 Free Will

Watson, G., ed., 2003. *Free Will*. 2nd edn, Oxford Readings in Philosophy. Oxford: OUP.

A collection of classic articles, including pieces by Chisholm, van Inwagen, Strawson (P. F.), Lewis, Frankfurt, Strawson (G.), Nagel, O'Connor, Kane, Scanlon, and Wolf.

Dennett, D., 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge, MA: MIT Press.

This is probably the most subtle defense of compatibilism available and is Dennett at his best.

Kane, R., ed., 2005. *The Oxford Handbook of Free Will*. Oxford: OUP.

An up-to-date and comprehensive anthology, containing articles by the leading contemporary libertarians and soft and hard determinists; there are useful sections on determinism and modern physics, and the impact of neuroscience.

Swinburne, R., ed., 2011. *Free Will and Modern Science*. Oxford: British Academy and OUP.

Both neurology and quantum theory are thought by some philosophers to throw light on the problem of free will, either for or against libertarianism. This collection has statements of all the major points of view on this question.

2.10 Knowledge

DeRose, K., 1995. "Solving the Skeptical Problem." *The Philosophical Review*, 104, pp. 1–52.

Stewart Cohen and DeRose developed the contextualist theory of knowledge attributions. DeRose's 1995 article has become a standard in the field.

Shope, R., 2002. "Conditions and Analyses of Knowing." In P. Moser, ed., *The Oxford Handbook of Epistemology*. Oxford: OUP, pp.25–71.

An overview of issues that arise in connection with the analysis of knowledge.

Sosa, E., 1991. *Knowledge in Perspective*. Cambridge: CUP.

Influential essays from a leading contemporary philosopher, on topics ranging from the Gettier problem to virtue epistemology.

Stroud, B., 1984. *The Significance of Philosophical Skepticism*. Oxford: OUP.

An influential treatise on skepticism but before the contemporary emphasis on knowledge attributions. A powerful defense of the skeptic's reasoning.

Zagzebski, L., 1996. *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Wellbeing Foundations of Knowledge*. Cambridge: CUP.

Groundbreaking work on intellectual virtue, articulating and addressing some basic questions about the value of knowledge.

2.11 Perception

Price, H. H., 1932. *Perception*. London: Methuen.

This is the classic statement of and argument for the sense-datum theory, and immediately precedes the positivist transmutation of that theory into a linguistic one: Price's sense-data are meant to be real entities and not merely the creatures of a "sense-datum language" as they are in Ayer's *Foundations of Empirical Knowledge*, for example.

Austin, J. L., 1962. *Sense and Sensibilia*. Oxford: Clarendon.

This book was constructed from lectures Austin gave in the 1950s. The received wisdom was that he conclusively refuted the sense-datum theory but, though this conviction remains quite strong, it is no longer taken so clearly for granted. Some philosophers feel that this work contains more sneering and polemic than straight argument. Ayer ended the lecture titled "Has Austin Refuted the Sense-datum Theory?" which he delivered in 1967, thus: "It is, in my view, a tribute to [Austin's] wit and to the strength of his personality that he was able to persuade so many philosophers—some of whom were quite able men!" The printed version omits the last phrase. Nevertheless, contemporary "relationist" direct realists often recur to Austin's argument and so the work retains the status of a classic.

Robinson, H., 1994. *Perception*. London: Routledge.

This book in "The Problems of Philosophy" series discusses theories in the philosophy of perception from the Greeks to the present day and defends the causal-hallucinatory argument against direct realism. This involves discussions of the early versions of disjunctivism.

Gendler, T. and Hawthorne, J., eds, 2006. *Perceptual Experience*. Oxford: OUP.

Probably the most comprehensive recent collection of articles that covers the issues that are to be found in the current philosophy of perception.

2.12 Ethics

Moore, G. E., *Principia Ethica*, for reasons given above.

Hare, R. M., 1952. *The Language of Morals*. Oxford: Clarendon; also

Hare, R. M., 1965. *Freedom and Reason*. Oxford: Clarendon.

Hare's significance consists mainly in the way he tries to make noncognitivist ethics also rationalist, by building up the principle of universalizability in an almost Kantian fashion.

Miller, A., 2003. *An Introduction to Contemporary Metaethics*. Oxford: Polity.

A succinct but wide-ranging guide to analytic metaethics, with chapters on intuitionism, emotivism, Blackburn's quasi-realism, Gibbard's norm-expressivism, "Cornell Realism," Railton's reductionism, and McDowell's realism.

Parfit, D., 2011. *On What Matters, Vols 1 and 2*. Oxford: Clarendon.

This contains 25 years of reflection on how to improve what he says in *Reasons and Persons*. It was a much-awaited text. Parfit attempts to develop a rationalist and objectivist ethics for a wholly secular world-view.

2.13 Political Philosophy

Rawls, J., 1971/1999. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

The importance of this book for political philosophy cannot be overemphasized. Before its publication the discipline was, in Peter Laslett's words, dead. This book aims to generalize the social contract theory, and defends two principles of justice that defend both freedom and equality.

Rawls, J., 1993/1996. *Political Liberalism*. New York: Columbia University Press.

In this book Rawls responds to criticisms his *Theory of Justice* received. In the former book he argues for a conception of justice that depends on strong metaphysical assumptions. In this book he argues that liberalism is best defended by appealing only to political value, setting aside controversial metaphysical, moral, or religious doctrines.

Nozick, R., 1974. *Anarchy, State and Utopia*. Oxford: Blackwell.

Usually regarded as the right-wing libertarian response to Rawls' interventionist liberalism, it earned its author the epithet "the right wing Rawls." It presents a variety of challenges to the legitimacy of state powers.

Cohen, G. A., 2008. *Rescuing Justice and Equality*. Cambridge, MA: Harvard University Press.

In this book Cohen launches a serious critique of Rawls' theory of justice. He argues that the methodology used by Rawls (political constructivism) is mistaken since principles of justice are insensitive to facts. More substantively, he defends two theses. First, he argues that justice requires more equality than Rawls permits. Second, he argues that in a just society, individuals' daily lives would be governed by the same principles that govern their institutions.

Dworkin, R., 2002. *Sovereign Virtue*. Cambridge, MA: Harvard University Press.

Dworkin presents an account of justice that incorporates personal responsibility into equality. He argues that treating people as equals requires that they are liable to bear the costs of their choices, but inequalities are unjustified when they are the product of bad brute luck.

Bibliography

(This bibliography is for the Preface, and Parts I and III)

- Albert, D. Z., 1992. *Quantum Mechanics and Experience*. Harvard: Harvard University Press.
- Allard, J. W., 2005. *The Logical Foundations of Bradley's Metaphysics: Judgment, Inference, and Truth*. Cambridge: CUP.
- Ambrose, A. and Lazerowitz, M., eds, 1970. *G.E. Moore: Essays in Retrospect*. London: George Allen and Unwin.
- Amoretti, M. and Vassallo, N., eds, 2009. *Knowledge, Language, and Interpretation: On the Philosophy of Donald Davidson*. Frankfurt-Heusenstamm: Ontos Verlag.
- Anscombe, G. E. M., 1958. "Modern Moral Philosophy." *Philosophy*, 33, pp. 1–19.
- Apel, K., 1988. *Diskurs und Verantwortung: Das Problem des Übergangs zur postkonventionellen*. Frankfurt: Suhrkamp.
- Armstrong, D., 1989. *A Combinatorial Theory of Possibility* Cambridge: CUP.
- , 1968. *A Materialist Theory of the Mind*. London: Routledge and Kegan Paul.
- Austin, J. L., 1962. *Sense and Sensibilia*. Oxford: Clarendon.
- , 1956–7. "A Plea for Excuses." *Proceedings of the Aristotelian Society*, 57, pp. 1–30.
- Ayer, A. J., 1946 (1971). *Language, Truth and Logic*. 2nd edn. Harmondsworth: Pelican.
- Bach, K., 1987. *Thought and Reference*. Oxford: OUP.
- Baker, G. P., 1988. *Wittgenstein: Meaning and Understanding: Essays on the Philosophical Investigations*. Oxford: Basil Blackwell.
- Baldwin, T., ed., 1993. *G.E. Moore: Selected Writings*. London: Routledge.
- Barrett, J. A., 2001. *The Quantum Mechanics of Minds and Worlds*. Oxford: OUP.
- Bayne, T., 2010. *The Unity of Consciousness*. Oxford: OUP.
- Bayne, T. and Montague, M., eds, 2011. *Cognitive Phenomenology*. Oxford: OUP.
- Beaney, M., ed., 2007. *The Analytic Turn: Analysis in Early Analytic Philosophy*. London: Routledge.
- Beckermann, A., 2004. "Einleitung." In P. Precht, ed., *Grundbegriffe der Analytischen Philosophie*. Stuttgart: Metzler, pp. 1–12.
- Bell, D., 1999. "The Revolution of Moore and Russell: A Very British Coup?" *Royal Institute of Philosophy Supplement*, 44, pp. 193–209.
- Bell, J. S., 1987. *Speakable and Unsayable in Quantum Mechanics*. Cambridge: CUP.
- Berger, A., ed., 2011. *Saul Kripke*. Cambridge: CUP.

- Black, M., 1944. The "Paradox of Analysis." *Mind*, 53(211), pp. 263–7.
- Blackburn, S., 1998. *Ruling Passions*. Oxford: Clarendon.
- , 1987. "Morals and Modals." In G. Macdonald and C. Wright, eds, *Fact, Science and Morality, Essays in Honour of A.J. Ayer's Language, Truth and Logic*. Oxford: Blackwell, pp. 119–41.
- , 1984. *Spreading the Word*. Oxford: OUP.
- Block, N., 1986. "Advertisement for a Semantics for Psychology." In P. A. French, T. E. Uehling, and H. K. Wettstein, eds, *Midwest Studies in Philosophy Vol. X*. Minneapolis: University of Minnesota Press, pp. 617–78.
- Bradley, F. H., 1935. *Collected Essays*. Oxford: Clarendon.
- , 1914. *Essays on Truth and Reality*. Oxford: Clarendon.
- , 1911. "On Some Aspects of Truth." *Mind*, 20(79), pp. 305–41.
- , 1893. *Appearance and Reality*. London: Swan Sonnenschein.
- Burge, T., 1979. "Individualism and the Mental." *Midwest Studies in Philosophy*, 4, pp. 73–121.
- Burgess, J., forthcoming. *Saul Kripke: Puzzles and Mysteries*. Key Contemporary Thinkers, Polity.
- Butterfield, J., 1988. "Albert Einstein meets David Lewis." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 2, pp. 65–81.
- Callender, C., 2011. "Philosophy of Science and Metaphysics." In S. French and S. Juha, eds, *The Continuum Companion to the Philosophy of Science*. London: Continuum, pp. 33–54.
- Cameron, R., 2007. "How to Have a Radically Minimal Ontology." *Philosophical Studies*, 151, pp. 249–264.
- Candlish, S., 2007. *The Russell/Bradley Dispute and its Significance for Twentieth-century Philosophy*. Basingstoke & New York: Palgrave.
- Carnap, R., 1969. *The Logical Structure of the World and Pseudoproblems in Philosophy*. Berkeley and Los Angeles: University of California Press.
- , 1950. "Empiricism, Semantics and Ontology." *Revue Internationale de Philosophie*, 4(2), pp. 20–40.
- , 1947. *Meaning and Necessity*. Chicago: University of Chicago Press.
- , 1942. *Introduction to Semantics*. Cambridge, MA: Harvard University Press.
- , 1937. *The Logical Syntax of Language*. London: Kegan Paul.
- , 1935. *Philosophy and Logical Syntax*. London: Kegan Paul.
- , 1932–3. "Psychology in a Physical Language." *Erkenntnis*, 3, pp. 107–42.
- , 1928. *Der logische Aufbau der Welt*. Berlin-Schlachtensee: Weltkreis-Verlag.
- Chalmers, D., 2012. *Constructing the World*. Oxford: OUP.
- , 2010. *The Character of Consciousness*. Oxford: OUP.
- , 2006. "The Foundations of Two-Dimensional Semantics." In M. García-Carpintero and J. Macià, eds, *Two-Dimensional Semantics: Foundations and Applications*. Oxford: OUP, pp. 55–140.
- , 1996. *The Conscious Mind*. Oxford: OUP.
- Chalmers, D., Manley, D. and Wasserman, R., eds, 2009. *Metametaphysics, New Essays on the Foundations of Ontology*. Oxford: OUP.
- Chomsky, N., 1959. "Review of Verbal Behavior." *Language*, 35, pp. 26–58.
- Churchland, P. M., 1981. "Eliminative Materialism and the Propositional Attitudes." *Journal of Philosophy*, 78, pp. 67–90.

- Coffa, J. A., 1991. *The Semantic Tradition from Kant to Carnap: To the Vienna Station*. Cambridge: CUP.
- Coleman, S., 2012. "Mental Chemistry: Combination for Panpsychists." *Dialectica*, 66(1), pp. 137–66.
- Collingwood, R. G., 1939. *An Autobiography*. Oxford: OUP.
- Crane, T., 1992. "The Nonconceptual Content of Experience." In T. Crane, ed., *The Contents of Experience*. Cambridge: CUP, pp. 136–57.
- Crick, F., 1994. *The Astonishing Hypothesis: The Scientific Search for the Soul*. New York: Scribners.
- Cussins, A., 1990. "The Connectionist Construction of Concepts." In M. Boden, ed., *The Philosophy of Artificial Intelligence*. Oxford: OUP, pp. 368–440.
- Dainton, B., 2012. "Selfhood and the Flow of Experience." *Grazer Philosophische Studien*, 84, pp. 161–99.
- , 2010. *Time and Space*. 2nd edn. Durham: Acumen.
- , 2008. *The Phenomenal Self*. Oxford: OUP.
- , 2006. *Stream of Consciousness*. 2nd edn. London: Routledge.
- Damasio, A., 1999. *The Feeling of What Happens: Body, Emotion and the Making of Consciousness*. Florida: Harcourt.
- Davidson, D., 2007. "Radical Interpretation." In *Dialectica* (reprinted in Davidson 2001b), pp. 314–28.
- , 2001c. *Subjective, Intersubjective, Objective*. Oxford: OUP.
- , 2001b. *Inquires into Truth and Interpretation*. Oxford: OUP.
- , 2001a. *Essays on Actions and Events*. Oxford: OUP.
- , 1990. "The Structure and Content of Truth (The Dewey Lectures 1989)." *Journal of Philosophy*, 87, pp. 279–328.
- , 1975. "Thought and Talk." In S. Guttenplan, ed., *Mind and Language*. Oxford: OUP, pp. 7–23 (reprinted in Davidson 2001b).
- , 1974. "On the Very Idea of a Conceptual Scheme." *Proceedings and Addresses of the American Philosophical Association*, 47, pp. 5–20 (reprinted in Davidson 2001b).
- , 1970. "Mental Events." In L. Foster and J. W. Swanson, eds, *Experience and Theory*. London: Duckworth, pp. 79–101 (reprinted in Davidson 2001a).
- , 1967. "Truth and Meaning." *Synthese*, 17, pp. 130–46 (reprinted in Davidson 2001b).
- , 1963. "Actions, Reasons and Causes." *Journal of Philosophy*, 60, pp. 685–700 (reprinted in Davidson 2001a).
- De Gaynesford, M., 2006. *Hilary Putnam*. Durham: Acumen.
- Dennett, D., 2012. "Sakes and Dints—And Other Definitions that Philosophers Really Need not Seek." *Times Literary Supplement*, March 2, pp. 12–14.
- , 1988. "Quining Qualia." In A. Marcel and E. Bisiach, eds, *Consciousness in Contemporary Science*. New York: OUP, pp. 42–77.
- Dorr, C. and Rosen, G., 2002. "Composition as a Fiction." In R. Gale, ed., *The Blackwell Guide to Metaphysics*. Oxford: Blackwell, pp. 151–74.
- Dreben, B. and Floyd, J., 2011. "Frege-Wittgenstein Correspondence." In E. D. Pellegrin, ed., *Interactive Wittgenstein: Essays in Memory of Georg Hendrik von Wright*. Dordrecht: Springer, pp. 15–73.
- Dretske, F., 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT.

- Dummett, M., 2010. *The Nature and Future of Philosophy*. New York: Columbia University Press.
- , 1993. *Origins of Analytic Philosophy*. London: Duckworth.
- , 1991. *Frege: Philosophy of Mathematics*. London: Duckworth.
- , 1983. *The Interpretation of Frege's Philosophy*. London: Duckworth.
- , 1978. *Truth and Other Enigmas*. London: Duckworth.
- , 1973. *Frege: Philosophy of Language*. London: Duckworth.
- Earman, J., 1995. *Bangs, Crunches, Whimpers and Shrieks: Singularities and Acausalities in Relativistic Spacetimes*. Oxford: OUP.
- , 1992. *World Enough and Space-Time*. Cambridge, MA: MIT.
- Earman, J. and Norton, J. D., 1987. "What Price Spacetime Substantivalism?" *British Journal for the Philosophy of Science*, 38, pp. 515–25.
- Ebbs, G., 2011. "Carnap and Quine on Truth by Convention." *Mind*, 120(478), pp. 193–237.
- Eco, U., 1996. *The Search for the Perfect Language*, Oxford, Blackwell.
- Edelman, G. M., 1992. *Bright Air, Brilliant Fire: On the Matter of the Mind*. New York: Basic Books.
- Edmonds, D. and Edinow, W., 2001. *Wittgenstein's Poker: The Story of a 10-minute argument between two great philosophers*. New York: Harper Collins.
- Evans, G., 1982. *The Varieties of Reference*. Oxford: OUP.
- , 1973. "The Causal Theory of Names." *Proceedings of the Aristotelian Society*, Supplementary Volume 47, pp. 187–208.
- Farkas, K., 2008. "Phenomenal Intentionality without Compromise." *The Monist*, 91, pp. 273–93.
- Feinberg, B. and Kasrils, R., eds, 1969. *Dear Bertrand Russell . . . A Selection of his Correspondence with the General Public 1950–1968*. London: George Allen and Unwin.
- Field, H., ed., 1980. *Science Without Numbers*. Princeton: Princeton University Press.
- Fine, K., 2000. "A Counter-example to Locke's Thesis." *The Monist*, 83, pp. 357–61.
- Fodor, J. A., 1987. *Psychosemantics*. Cambridge, MA: Bradford Books.
- , 1981. *Representations*. Cambridge, MA: MIT.
- , 1975. *The Language of Thought*. New York: Thomas Crowell.
- Fodor, J. A. and Lepore, E., 1994. "What is the Connection Principle?" *Philosophy and Phenomenological Research*, 54, pp. 837–45.
- Foot, P., 1958–9. "Moral Beliefs." *Proceedings of the Aristotelian Society*, 59, pp. 83–104.
- , 1958. "Moral Arguments." *Mind*, 67(268), pp. 502–13.
- Foster, J., 1991. *The Immaterial Self*. London: Routledge.
- , 1985. *Ayer*. London: Routledge.
- , 1982. *The Case for Idealism*. London: Routledge and Kegan Paul.
- Freeman, A., ed., 2006. *Consciousness and its Place in Nature: Does Physicalism Entail Panpsychism?* Exeter: Imprint Academic.
- Frege, G., 1980. *Philosophical and Mathematical Correspondence*. Oxford: Basil Blackwell.
- , 1964. [Grundgesetze der Arithmetik.] *The Basic Laws of Arithmetic. Exposition of the System*. Translated and edited, with an introduction, by Montgomery Furth. University of California Press: Berkeley & Los Angeles.
- , 1956. "The Thought: A Logical Inquiry." *Mind*, 65(259), pp. 289–311.

- , 1953. *The Foundations of Arithmetic*. Second revised edition, translation by J. L. Austin. German and English. Oxford: Basil Blackwell.
- , 1952. *Translations from the Philosophical Writings of Gottlob Frege*. Edited by Peter Geach and Max Black. Oxford: Basil Blackwell.
- , 1879/1967. *Begriffsschrift. A Formula Language, Modeled upon that of Arithmetic, for Pure Thought*. Translated by Stefan Bauer-Mengelberg, in J. van Heijenoort (ed.) *From Frege to Gödel: A Sourcebook in Mathematical Logic, 1879–1931*. Cambridge MA: Harvard University Press, pp. 1–82.
- Friedman, M., 2000. *A Parting of the Ways: Carnap, Cassirer, and Heidegger*. Chicago: Open Court.
- , 1999. *Reconsidering Logical Positivism*. Cambridge: CUP.
- Gabriel, G., ed., 1980. *Frege Philosophical and Mathematical Correspondence*. Chicago: University of Chicago Press.
- Galison, P., 1990. "Aufbau/Bauhaus: Logical Positivism and Architectural Modernism." *Critical Inquiry*, 16, pp. 709–52.
- Gaskin, R., 2010. "Précis of The Unity of the Proposition." *Dialectica*, 64(2), pp. 259–64.
- , 2008. *The Unity of the Proposition*. Oxford: OUP.
- Gaskin, R., ed., 2001. *Grammar in Early Twentieth-Century Philosophy*. London: Routledge.
- Gendler, T. and Hawthorne, J., eds, 2006. *Perceptual Experience*. Oxford: OUP.
- George, A. and Heck, R., 1998. "Frege." In E. Craig, ed., *Routledge Encyclopedia of Philosophy*. London: Routledge, pp. 765–78.
- Gibbard, A., 1990. *Wise Choices, Apt Feelings*. Oxford: Clarendon.
- Goff, P., 2011. "There is no Combination Problem." In M. Blamauer, ed., *The Mental as Fundamental: New Perspectives on Panpsychism*. Frankfurt: Ontosverlag, pp. 131–40.
- Goldstein, L., 1999. "Wittgenstein's Ph.D Viva—A Re-creation." *Philosophy*, 74, pp. 499–513.
- Goodman, N., 1963. "The Significance of *Der Logische Aufbau Der Welt*." In P. A. Schilpp, ed., *The Philosophy of Rudolf Carnap*. La Salle, IL: Open Court, pp. 545–8.
- , 1951. *The Structure of Appearance*. Cambridge, MA: Harvard University Press.
- Grattan-Guinness, I., ed., 1977. *Dear Russell – Dear Jourdain*. London: Duckworth.
- Greene, B., 2004. *The Fabric of the Cosmos*. New York: Knopf.
- Grice, P., 1989. *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
- , 1975. "Logic and Conversation." In D. Davidson and G. Harman, eds, *The Logic of Grammar*. Encino, CA: Dickenson, pp. 64–75.
- , 1957. "Meaning." *The Philosophical Review*, 66, pp. 377–88.
- Grice, P. and Strawson, P. F., 1957. "In Defence of a Dogma." *The Philosophical Review*, 65, pp. 141–58.
- Griffin, N., 1991. *Russell's Idealist Apprenticeship*. Oxford: Clarendon.
- Griffin, N., ed., 1992. *The Selected Letters of Bertrand Russell: The Private Years 1884–1914*. London: Penguin.
- Gutting, G., 2011. *Thinking the Impossible: French Philosophy Since 1960*. Oxford and New York: OUP.

- Habermas, J., 1998. *Between Facts and Norms. Contributions to a Discourse Theory of Law and Democracy*. Cambridge, MA: MIT.
- Hacker, P. M. S., 1998. "Analytic Philosophy: What, Whence and Whither?" In A. Biletzki and A. Matar, eds, *The Story of Analytic Philosophy—Plot and Heroes*. London: Routledge, pp. 3–34.
- , 1996b. *Wittgenstein's Place in Twentieth-century Analytic Philosophy*. Oxford: Blackwell.
- , 1996a. *An Analytical Commentary on the Philosophical Investigations*. Vol. 4, *Wittgenstein: Mind and Will*. Oxford: Blackwell.
- , 1993. *Analytic Commentary on the Philosophical Investigations*. Vol. 3, *Wittgenstein Meaning and Mind*. Part 2, *Exegesis [para.]* 243–427. Oxford: Blackwell.
- , 1990. *An Analytic Commentary on the "Philosophical Investigations."* Vol. 3, *Wittgenstein : Meaning and Mind*. Oxford: Blackwell.
- , 1975. *Insight and Illusion : Wittgenstein on Philosophy and the Metaphysics of Experience*. London: OUP.
- Hacker, P. M. S. and Baker, G. P., 1985. *Wittgenstein: Rules, Grammar and Necessity, Volume 2*. Oxford: Blackwell.
- , 1980. *Wittgenstein: Understanding and Meaning, Volume 1 of an Analytical Commentary on the Philosophical Investigations*. Oxford: Blackwell.
- Hacker, P. M. S. and Bennett, M., 2003. *Philosophical Foundations of Neuroscience*. Oxford: Blackwell.
- Hahn, L. E., ed., 1999. *The Philosophy of Donald Davidson, Library of Living Philosophers v. 27*. Chicago: Open Court.
- Hare, R., 1971. "Wanting: Some Pitfalls." In R. Branaugh, R. Williams, and A. Marras, eds, *Agent, Action and Reason*. Toronto: Toronto University Press, pp. 81–97.
- , 1964. "Pain and Evil." *Proceedings of the Aristotelian Society*, 91, pp. 91–106.
- , 1963. *Freedom and Reason*. London: OUP.
- , 1952. *The Language of Morals*. London: OUP.
- Harman, G., 1990. "The Intrinsic Quality of Experience." In N. Block, ed., *The Nature of Consciousness*. Cambridge, MA: MIT, pp. 31–52.
- Hellman, G., 1981. "How to Gödel a Frege-Russell: Gödel's Incompleteness Theorems and Logicism." *Noûs*, 15(4), pp. 451–68.
- Hempel, C., 1949. "The Logical Analysis of Psychology." In H. Feigl and W. Sellars, eds, *Readings in Philosophical Analysis*. New York: Appleton-Century-Crofts, pp. 373–84.
- Horgan, T. and Tienson, J., 2002. "The Intentionality of Phenomenology and the Phenomenology of Intentionality." In D. Chalmers, ed., *Philosophy of Mind: Classical and Contemporary Readings*. Oxford: OUP, pp. 520–33.
- Horgan, T., Tienson, J., and Graham, G., 2004. "Phenomenal Intentionality and the Brain in a Vat." In R. Scantz, ed., *The Externalist Challenge: New Studies on Cognition and Intentionality*. Amsterdam: de Gruyter, pp. 41–62.
- Hylton, P., 2007. *Quine*. New York: Routledge.
- , 1990. *Russell, Idealism, and the Emergence of Analytic Philosophy*. Oxford: Clarendon.
- Jackson, F., 1982. "Epiphenomenal Qualia." *Philosophical Quarterly*, 82, pp. 127–36.
- Jiang, Y. and Bai, T., 2010. "Studies in Analytic Philosophy in China." *Synthese*, 175, pp. 3–12.

- Johnston, M., 2010. *Surviving Death*. Princeton: Princeton University Press.
- Kenny, A., 2006. *Wittgenstein*. 2nd edn. Oxford: Wiley-Blackwell.
- , 1984. *The Legacy of Wittgenstein*. Oxford: Blackwell.
- Keynes, J. M., 1938. "My Early Beliefs." In J. M. Keynes, ed., *The Collected Writings of John Maynard Keynes*, Vol. 10. London: Macmillan, pp. 433–50.
- Kind, A., 2010. "Transparency and Representationalist Theories of Consciousness." *Philosophy Compass*, 5(10), pp. 902–13.
- Knobe, J. and Nichols, S., eds, 2008. *Experimental Philosophy*. Oxford: OUP.
- Kriegel, U., 2007. "Intentional Inexistence and Phenomenal Intentionality." *Philosophical Perspectives*, 141, pp. 307–40.
- Kripke, S., 1982. *Wittgenstein on Rules and Private Language*. Oxford: Blackwell.
- , 1980. *Naming and Necessity*. Oxford: Basil Blackwell.
- Kuhn, T., 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- Leiter, B., ed., 2004. *The Future of Philosophy*. Oxford: OUP.
- Leitgeb, H., (forthcoming). "New Life for Carnap's *Aufbau*?" *Synthese*.
- Lepore, E. and Ludwig, K., 2007. *Donald Davidson's Truth-Theoretic Semantics*. Oxford: Clarendon.
- , 2005. *Donald Davidson: Meaning, Truth and Language*. Oxford: Clarendon.
- Levine, J., 1983. "Materialism and Qualia: The Explanatory Gap." *Pacific Philosophical Quarterly*, 64, pp. 354–61.
- Levy, P., 1981. *Moore : G.E. Moore and the Cambridge Apostles*. Oxford: OUP.
- Lewis, D., 1986b. *Philosophical Papers*, Volume 2. Oxford: OUP, pp. ix–x.
- , 1986a. *On the Plurality of Worlds*. Oxford: Blackwell.
- , 1972. "Psychophysical and Theoretical Identifications." *Australasian Journal of Philosophy*, 50(3), pp. 249–58.
- Linsky, B. and Zalta, E. N., 2006. "What is Neologicism?" *The Bulletin of Symbolic Logic*, 12(1), pp. 60–99.
- Loar, B., 2003. "Phenomenal Intentionality as the Basis of Mental Content." In H. Martin and B. Ramberg, eds, *Reflections and Replies: Essays on the Philosophy of Tyler Burge*. Cambridge: MIT Press, pp. 229–58.
- , 1987. "Subjective Intentionality." *Philosophical Topics*, 15, pp. 89–124.
- Lockwood, M., 1989. *Mind, Brain and Quantum*. Oxford: OUP.
- Loewer, B., 2004. "'David Lewis' Humean Theory of Objective Chance," *Philosophy of Science* 71(5), pp. 1115–25.
- Luckhardt, C. G., ed., 1979. *Wittgenstein: Sources and Perspectives*. Ithaca: Cornell University Press.
- Mabe, M., 2003. "The Growth and Number of Journals." *Serials: the Journal for the Serials Community*, 16(2), pp. 191–7.
- Mackie, J. L., 1977. *Ethics Inventing Right and Wrong*. London: Penguin.
- Magee, B., ed., 1971. *Modern British Philosophy*. London: Secker and Warburg.
- Marr, D., 1983. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: W.H. Freeman and Company.
- Maudlin, T., 2004. "Thoroughly Muddled McTaggart: Or, How to Abuse Gauge Freedom to Create Metaphysical Monstrosities." *Philosophers Imprint*, 2(3), pp. 1–23.

- Maudlin, T., 1990. "Substances and Spacetimes: What Aristotle Would have Said to Einstein." *Studies in the History and Philosophy of Science*, 21, pp. 531–61.
- McDowell, J. H., 1994a. "The Content of Perceptual Experience." *Philosophical Quarterly*, 44, pp. 190–205.
- , 1994b. *Mind and World*. Cambridge, MA; London: Harvard University Press.
- , 1986. "Singular Thought and the Extent of Inner Space." In J. McDowell and P. Pettit, eds, *Subject, Thought and Context*. Oxford: Blackwell, pp. 137–68.
- McGinn, C., 2012. *Truth by Analysis*. Oxford: OUP.
- , 1977. "Charity, Interpretation and Belief." *Journal of Philosophy*, 74, pp. 521–35.
- McGuinness, B. F., ed., 1967. *Wittgenstein und der Wiener Kreis*. Oxford: Blackwell.
- Macpherson, F., ed., 2011. *The Senses: Classical and Contemporary Perspectives*. Oxford: OUP.
- Monk, R., 1996. *Bertrand Russell: The Spirit of Solitude*. London: Random House.
- , 1990. *Ludwig Wittgenstein: The Duty of Genius*. London: Jonathan Cape.
- Moore, A. W., ed., 1993. *Meaning and Reference*. Oxford: OUP.
- Moore, G. E., 1959. *Philosophical Papers*. London: George Allen and Unwin.
- , 1939. "Proof of an External World." *Proceedings of the British Academy*, 25, pp. 273–300.
- , 1903. *Principia Ethica*. Cambridge: CUP.
- , 1901–2. "Truth." In J. Baldwin, ed., *Dictionary of Philosophy and Psychology*. Reprinted in G.E. Moore: *Selected Writings*. London: Macmillan, pp. 20–2.
- , 1899. "The Nature of Judgment." *Mind*, 8, pp. 176–93.
- Morris, M., 2008. *Wittgenstein and the Tractatus Logico-Philosophicus*. Abingdon and New York: Routledge.
- Mumford, S., ed., 2003. *Russell on Metaphysics: Selections from the Writings of Bertrand Russell*. London: Routledge.
- Nagel, E., 1961. *The Structure of Science*. London: Routledge and Kegan Paul.
- Nagel, T., 1986. *The View from Nowhere*. Oxford: OUP, pp. 90–109.
- , 1974. "What is it Like to be a Bat?" *Philosophical Review*, LXXXIII, pp. 435–50.
- Nerlich, G., 1994. *The Shape of Space*. Cambridge: CUP.
- Neurath, O., 1973. *Empiricism and Sociology*. Dordrecht: D. Reidel.
- Norton, J., 1988. "The Hole Argument." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 2, pp. 56–64.
- Nozick, R., 1974. *Anarchy, State, and Utopia*. New York: Basic Books.
- Parfit, D., 2011. *On What Matters (Volumes One and Two)*. Oxford: OUP.
- Paul, L. A., 2010. "A New Role for Experimental Work in Metaphysics." *European Review of Philosophy and Psychology*, 1, pp. 461–76.
- Peacocke, C., 1992. *A Study of Concepts*. Cambridge, MA: MIT.
- , 1989. "Perceptual Content." In J. Almog, J. Perry, and H. Wettstein, eds, *Themes from Kaplan*. New York: OUP, pp. 297–329.
- , 1986. "Analogue Content." *Proceedings of the Aristotelian Society*, 60, pp. 1–17.
- Penrose, R., 2004. *The Road to Reality*. London: Jonathan Cape.
- , 1989. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford: OUP.
- Perez, D. I. and Ortiz-Milan, G., 2010. "Analytic Philosophy." In S. Niccetelli and O. Schutte, eds, *A Companion to Latin American Philosophy*. Oxford: Blackwell, pp. 199–214.

- Perry, J., 2001. *Reference and Reflexivity*. Stanford: CSLI Publications.
- Pessin, A. and Goldberg, S., eds, 1996. *The Twin Earth Chronicles: Twenty Years of Reflection on Hilary Putnam's "The Meaning of 'Meaning'."* New York: Sharpe.
- Pincock, C., 2009. "Carnap's Logical Structure of the World." *Philosophy Compass*, 4(6), pp. 951–61.
- Pitt, D., 2008. "Intentional Psychologism." *Philosophical Studies*, 146, pp. 117–38.
- , 2004. "The Phenomenology of Cognition; or *What Is It Like to Think that P?*" *Philosophy and Phenomenological Research*, 69, pp. 1–36.
- Plantinga, A., 2003. *Essays in the Metaphysics of Modality*. Oxford: OUP.
- Platts, M., 1979. *Ways of Meaning*. London: Routledge and Kegan Paul.
- Popper, K., 1959. *The Logic of Scientific Discovery*. London: Hutchinson.
- Price, H., 1996. *Time's Arrow and Archimedes Point*. Oxford: OUP.
- , 1989. "A Point on the Arrow of Time." *Nature*, 20(July 20), pp. 181–2.
- Prinz, J., 2007. *The Emotional Construction of Morals*. Oxford: OUP.
- Proops, I., 2006. "Russell's Reasons for Logicism." *Journal of the History of Philosophy*, 44, pp. 267–92.
- Putnam, H., 1981. *Reason, Truth and History*. Cambridge: CUP.
- , 1975. "The Meaning of 'Meaning.'" *Mind, Language and Reality: Philosophical Papers, Volume 2*. Cambridge: CUP, pp. 215–71.
- , 1967. "The Nature of Mental States." In W. H. Capitan and D. D. Merrill, eds, *Art, Mind, and Religion*. Pittsburgh: Pittsburgh University Press, pp. 37–48.
- , 1960. "Minds and Machines." In S. Hook, ed., *Dimensions of Mind*. New York: New York University Press, pp. 148–79.
- Quine, W. V. O., 1995. *From Stimulus to Science*. Cambridge, MA; London: Harvard University Press.
- , 1991. "Two Dogmas in Retrospect." *Canadian Journal of Philosophy*, 21(3), pp. 265–74.
- , 1990. *Pursuit of Truth*. Cambridge, MA/London: Harvard University Press.
- , 1985. *The Time of my Life : An Autobiography*. Cambridge, MA: MIT.
- , 1981. *Theories and Things*. Cambridge, MA: The Belknap Press of Harvard University Press.
- , 1974. *The Roots of Reference*. La Salle, IL: Open Court.
- , 1970. *The Web of Belief*. New York: Random House.
- , 1969b. "Reply to Chomsky." In D. Davidson and J. Hintikka, eds, *Words and Objections*. Dordrecht: D. Reidel, pp. 302–11.
- , 1969a. *Ontological Relativity and Other Essays*. New York: Columbia University Press.
- , 1960. *Word and Object*. New York: Technology Press of the Massachusetts Institute of Technology.
- , 1953. *From a Logical Point of View*. Cambridge, MA: Harvard University Press.
- , 1951. "Two Dogmas of Empiricism." *Philosophical Review*, 60, pp. 20–43.
- , 1936. "Truth by Convention," in O.H. Lee (ed.), *Philosophical Essays for A.N. Whitehead*. New York: Longmans.
- Ramachandran, M., 2008. "Descriptions and Presuppositions: Strawson vs. Russell." *South African Journal of Philosophy*, 27(3), pp. 64–79.
- Redhead, M. L. G., 1987. *Incompleteness, Nonlocality and Realism*. Oxford: Clarendon.

- Richardson, A., 1998. *Carnap's Construction of the World: the Aufbau and the Emergence of Logical Empiricism*. Cambridge: CUP.
- Richardson, A. and Uebel, T., eds, 2007. *The Cambridge Companion to Logical Empiricism*. Cambridge: CUP.
- Rickles, D., French, S., and Saatsi, J., eds, 2006. *The Structural Foundations of Quantum Gravity*. Oxford: OUP.
- Robinson, H., 2002. "Davidson and Non-reductive Physicalism: A Tale of Two Cultures." In B. Loewer and C. Gillett, eds, *Physicalism and Its Discontents*. Cambridge: CUP, pp. 129–51.
- , 1994. *Perception*. London: Routledge.
- , 1982. *Matter and Sense: A Critique of Contemporary Materialism*. Cambridge: CUP.
- Rorty, R., ed., 1967. *The Linguistic Turn*. Chicago: University of Chicago Press.
- Rosen, G., 1990. "Modal Fictionalism." *Mind*, 99, pp. 327–54.
- Russell, B., 2009. *Autobiography*. Routledge Classics edn. Abingdon: Routledge.
- , 1998. *Autobiography*. London: Routledge.
- , 1959/1985. *My Philosophical Development*. London: Unwin Hyman.
- , 1956. "The Philosophy of Logical Atomism." *Logic and Knowledge*. London: George Allen and Unwin, pp. 177–281.
- , 1951. *The Autobiography of Bertrand Russell*. London: George Allen and Unwin.
- , 1937. *The Principles of Mathematics*. 2nd edn. London: George Allen and Unwin.
- , 1927 (1954). *The Analysis of Matter*. London (New York): Kegan Paul (Dover).
- , 1918b. "The Relation of Sense-Data to Physics." *Mysticism and Logic*. London: George Allen and Unwin, pp. 145–79.
- , 1918a. *Mysticism and Logic*. London: Longmans Green.
- , 1915. "The Ultimate Constituents of Matter." *The Monist*, 25, pp. 399–417.
- , 1912. *The Problems of Philosophy*. London: Williams and Norgate.
- , 1910. "Knowledge by Acquaintance and Knowledge by Description." *Proceedings of the Aristotelian Society*, 11, pp. 108–28.
- , 1905. "On Denoting." *Mind*, 14, pp. 479–93.
- , 1903. *The Principles of Mathematics*. Cambridge: CUP.
- , 1901. "Recent Work on the Principles of Mathematics." *International Monthly*, 4, pp. 83–101.
- , 1894/2003. "On the Distinction Between the Psychological and Metaphysical Points of View." In S. Mumford, ed., *Russell on Metaphysics: Selections from the Writings of Bertrand Russell*. London: Routledge, pp. 21–4.
- Russell, B. and Whitehead, A. N., 1910–13. *Principia Mathematica*. Cambridge: CUP.
- Ryle, G., 1971. *Collected Papers*, Vol. 1. London: Hutchinson.
- , 1949. *The Concept of Mind*. London: Hutchinson; republished, Penguin, 1978.
- , 1931–2. "Systematically Misleading Impressions." *Proceedings of the Aristotelian Society*, 32, 139–70.
- Sainsbury, M., 2005. *Reference without Referents*. Oxford: OUP.
- Salmon, N., 1986. *Frege's Puzzle*. 2nd edn. Atacadero, California: Ridgeview.
- Saunders, S., Barrett, J., Kent, A., and Wallace, D., eds, 2010. *Many Worlds? Quantum Theory and Reality*. Oxford: OUP.
- Scanlon, T. N., 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

- Schaffer, J., 2010. "The Internal Relatedness of All Things." *Mind*, 119(474), pp. 341–76.
- Schilpp, P. A., ed., 1963. *The Philosophy of Rudolf Carnap*. Library of Living Philosophers, Volume XI. La Salle: Open Court.
- , 1944, *The Philosophy of Bertrand Russell*. Library of the Living Philosophers, Volume V. Chicago: Open Court.
- Schlick, M., 1979. "The Future of Philosophy." In H. L. Mulder and Van De Velde-Schlick, eds, *Philosophical Papers, Vol.2*. Dordrecht: Reidel, pp. 210–24.
- , 1959. "The Turning Point in Philosophy," translated by David Rynin. In A. J. Ayer, ed., *Logical Positivism*. Glencoe, IL: The Free Press, pp. 53–9.
- , 1939. *Problems of Ethics*. New York: Prentice-Hall.
- Searle, J., 2003. "Contemporary Philosophy in the United States." In N. Bunnin and E. P. Tsui-James, eds, *The Blackwell Companion to Philosophy*. 2nd edn. Oxford: Blackwell, pp. 1–22.
- , 1992. *The Rediscovery of Mind*. Cambridge, MA: MIT.
- , 1987. "Indeterminacy, Empiricism, and the First Person." *The Journal of Philosophy*, 84, pp. 123–46.
- , 1980. "Minds, Brains and Programs." *Behavioural and Brain Sciences*, 3, pp. 417–24.
- Sider, T., forthcoming. *Against Parthood*.
- , 2011. *Writing the Book of the World*. Oxford: OUP.
- Siewert, C. P., 1998. *The Significance of Consciousness*. Princeton, NJ: Princeton University Press.
- Simons, P., 2010. "Relations and Truthmaking." *Proceedings of the Aristotelian Society*, 84, pp. 199–213.
- , 1999. "Bolzano, Brentano and Meinong: Three Austrian Realists." *Royal Institute of Philosophy Supplement: German Philosophy Since Kant*, 44, pp. 109–36.
- Skinner, B. F., 1953. *Science and Human Behavior*. New York: Macmillan.
- Sklar, L., 1993. *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics*. Cambridge: CUP.
- Smart, J. J. C., 1959. "Sensations and Brain Processes." *Philosophical Review*, LXVIII, pp. 141–56.
- Soames, S., 2003b. *Philosophical Analysis in the Twentieth Century. Volume II: The Age of Meaning*. Princeton, NJ/Oxford: Princeton University Press.
- , 2003a. *Philosophical Analysis in the Twentieth Century. Volume I: The Dawn of Analysis*. Princeton, NJ/Woodstock: Princeton University Press.
- , 2002. *Beyond Rigidity*. Oxford: OUP.
- Sokal, A., 2008. *Beyond the Hoax: Science, Philosophy and Culture*. Oxford: OUP.
- Spadoni, C., 1978. "Philosophy in Russell's Letters to Alys." *Russell: the Journal of Bertrand Russell Studies*, 98(1), pp. 17–31.
- Sprigge, T. L. S., 1993. *James and Bradley: American Truth and British Reality*. Chicago and La Salle: Open Court.
- Stein, H., 1998. "Logicism." In E. Craig, ed., *Routledge Encyclopedia of Philosophy*. London: Routledge, pp. 811–17.
- Stoljar, D., 2001. "Two Conceptions of the Physical." *Philosophy and Phenomenological Research*, LXII(2), pp. 253–81.
- Strawson, G., 2009. *Selves*. Oxford: OUP.

- , 2005. "Real Intentionality 2." *Synthesis Philosophica*, 40, pp. 279–7.
- , 2004. "Real Intentionality." *Phenomenology and the Cognitive Sciences*, 3, pp. 287–313.
- Strawson, P. F., 1992. *Analysis and Metaphysics: An Introduction to Philosophy*. Oxford: OUP.
- , 1966. *The Bounds of Sense*. London: Methuen.
- , 1964. "Identifying Reference and Truth-Values." *Theoria*, 30, pp. 96–118.
- , 1959. *Individuals*. London: Methuen.
- , 1952. *Introduction to Logical Theory*. London: Methuen.
- , 1950. "On Referring." *Mind*, 59, pp. 320–44.
- , 1949. "Truth." *Analysis*, 9(6), pp. 83–97.
- Suits, B., 1978. *The Grasshopper: Games, Life and Utopia*. Toronto: University of Toronto Press.
- Textor, M., 2006. *The Austrian Contribution to Analytic Philosophy*. London and New York: Routledge.
- Tye, M., 2003. *Consciousness and Persons: Unity and Identity*. Cambridge, MA: MIT.
- , 2000. *Consciousness, Colour and Content*. Cambridge, MA: MIT.
- , 1995. *Ten Problems of Consciousness*. Cambridge, MA: MIT.
- Urmson, J. O., 1956. *Philosophical Analysis: Its Development between the Two World Wars*. Oxford: OUP.
- Valberg, J. J., 2007. *Dream, Death and the Self*. Princeton: Princeton University Press.
- Van Inwagen, P., 1990. *Material Beings*. Ithaca: Cornell University Press.
- Warnock, G. J., 1976. "Gilbert Ryle's Editorship." *Mind*, 85 (337), pp. 47–56.
- Williams, B., 1985. *Ethics and the Limits of Philosophy*. London: Fontana.
- , 1981. *Moral Luck*. Cambridge: CUP.
- Williamson, T., 2007. *The Philosophy of Philosophy*. Oxford: Blackwell.
- Wittgenstein, L., 1961. *TractatusLogico-Philosophicus*. D. F. Pears and B. F. McGuiness (trans.). New York: Humanities Press.
- , 1958. *The Blue and Brown Books*. Oxford: Blackwell.
- , 1953. *Philosophical Investigations*. G. E. M. Anscombe and R. Rhees (eds.), G. E. M. Anscombe (trans.). Oxford: Wiley-Blackwell.
- , 1929. "Some Remarks on Logical Form." *Proceedings of the Aristotelian Society*, 9, pp. 162–71.
- , 1922. *TractatusLogico-Philosophicus*. C.K. Ogden (trans.). London: Routledge and Kegan Paul.
- Wright, C., 1983. *Frege's Conception of Number as Objects*. Aberdeen: Aberdeen University Press.
- Zimmerman, D., 2004. "Metaphysics after the Twentieth Century." In D. Zimmerman, ed., *Oxford Studies in Metaphysics*, Volume I. Oxford: OUP, pp. ix–xxii.

Author Index

- Abizadeh, A. 545
Adler, A. 51
Aharonov, Y. 309
Albert, D. 293–6, 307–9, 553
Alston, W. 374
Amoretti, M. 157
Anderson, E. 496, 529
Anscombe, G. E. M. 83, 101–5, 157,
312, 318, 324, 468, 570, 584, 633,
635, 640
Apel, K. 157
Aristotle 19, 27, 75, 91, 103, 170, 285,
349, 353, 412, 448–9, 572–4, 611,
614, 623
Armstrong, C. 546
Armstrong, D. M. 131–2, 167, 266,
317–18, 371–3, 376, 452, 571, 635, 641
Arneson, R. 528
Aronson, J. 322
Austin, J. L. 83–6, 89–90, 98, 100,
460, 472, 570, 573, 584–5, 623,
634, 637, 644
Ayer, A. J. 53–7, 71, 86, 99, 102, 153,
265, 425, 440, 570, 577, 583–4, 593,
597, 629, 633–4, 644
Ayers, M. 412
Azzouni, J. 191

Bach, K. 85, 158, 216–23, 225
Bacon, F. 119, 256–8, 276
Baker, A. 191
Baker, G. 564, 633
Balashov, Y. 275
Barbour, J. 307
Baril, A. 447

Barrett, J. 553
Batterman, R. 278, 639
Bayne, T. 557, 579
Beaney, M. 561, 631, 635
Beckermann, A. 578
Beebee, H. 134, 167, 316, 318, 322, 640
Bell, C. 150
Bell, J. 305, 308–9, 581
Bell, V. 150
Benacerraf, P. 163, 185–7, 194, 196–8, 636
Bennett, J. 328
Bennett, K. 331, 579
Bentham, J. 15, 501
Berger, A. 158
Bergmann, G. 50
Berkeley, G. 5, 91, 176–7, 198, 267, 457,
572, 614, 622
Bernoulli, J. 259
Biletzki, A. 563
Bird, A. 275–8, 638
Black, M. 88
Blackburn, S. 56, 106, 153, 160, 237,
251–2, 571, 577, 581, 594, 645
Blatti, S. 379–80
Boghossian, P. 242, 247, 251–2
Bohm, D. 303–5, 308–9
Bohr, N. 303
Boltzmann, L. 50–1, 152, 166, 292–3,
302, 307
Bolzano, B. 51, 176, 198, 570, 581–2, 635
Boole, G. 25, 172
Boolos, G. 187, 198
Bosanquet, B. 3, 574
Boudry, M. 276
Bovens, L. 276

- Boyd, R. 268, 270, 275, 277–8
Bradley, F. 3–11, 13, 19, 148–9, 564, 574,
580, 628, 632
Braithwaite, R. 265
Brandom, R. 252, 585
Brandt, R. 484
Breckendridge, W. 472
Brentano, F. 51, 362, 570, 581, 603
Broad, C. D. 583, 635, 641
Bromberger, S. 227
Broome, J. 496
Brouwer, L. E. J. 163, 178–80, 182,
604, 636
Brown, H. 275–6
Brush, S. G. 278
Buckwalter, W. 443
Burke, M. 346
Byrne, A. 374, 376, 472

Cahn, S. 419
Callender, C. 271, 278, 307–8, 560–1
Campbell, C. A. 558
Campbell, J. 472
Candlish, S. 149, 561, 632
Caney, S. 544, 546
Cantor, G. 18–19, 23, 176–80, 183, 570,
581, 592–3
Carnap, R. 37, 50, 51–3, 56–60, 71–2,
77, 87, 100, 124, 153–5, 256, 261, 263,
265, 345, 553, 561–3, 566–7, 570,
573–4, 579, 581, 583, 590, 600, 608,
617, 633–6
Carroll, S. 308
Carston, R. 221, 223
Cartwright, N. 262, 266–7, 271
Cassam, Q. 472
Cauchy, A. L. 18, 176–7
Chakravartty, A. 277
Chalmers, D. 133, 135–6, 159, 168, 358,
360, 375, 561–2, 566–7, 581, 585, 627,
641–2
Chang, R. 107, 170, 494–6
Child, W. 472
Chisholm, R. 343, 557, 635, 642
Chomsky, N. 132, 164, 204–5, 227,
584, 635
Chrisman, M. 448

Christiano, T. 531, 536–7
Churchland, P. M. 134, 261, 268, 277–8
Clarke, R. 416
Clarke, S. 484
Clayton, M. 538
Cleland, C. 278
Cohen, G. A. 171, 529–30, 538–44,
547–8, 645
Cohen, S. 437, 643
Coleman, S. 554
Collins, J. 227, 639
Colyvan, M. 191–2, 637
Compte, A. 256
Conee, E. 440
Cover, J. 275, 277
Craig, E. 439
Craig, W. L. 419
Crane, T. 133
Craver, C. 324
Crick, F. 159
Curd, M. 275, 277
Cussins, A. 133

Dainton, B. 252, 375, 555, 560, 579, 638
Damasio, A. 159
Dancy, J. 484, 494
Darden, L. 324
Darwall, S. 106, 484–5
Darwin, C. 15, 75, 256, 274, 449
Davidson, D. 108–17, 131, 134, 157–8,
207–9, 328, 369, 472, 495, 563–4, 570,
573, 577, 584–5, 591, 608, 624, 637
Dawes Hicks, G. 558
de Fineti, B. 260
Dedekind, R. 18–19, 23, 174, 176–7,
570, 581
Deleuze, G. 576
Democritus 285
Dennett, D. 134, 157, 428, 575, 585,
641, 643
DeRose, K. 436–8, 643
Derrida, J. 576–7, 585
Descartes, R. 36, 87, 91, 93, 116, 160,
256, 258, 264, 267, 306, 357–8, 572
Devitt, M. 227
Dick, M. 156
Dogramaci, S. 439

- Donnellan, K. 158
Dorsey, D. 509
Douven, I. 276
Dowe, P. 167, 322–3, 332
Dreben, B. 152
Dretske, F. 133, 266, 435, 448, 469, 472
Ducasse, C. J. 558
Duhem, P. 51, 255, 262, 267–8, 277
Dummett, M. 25, 202, 205, 306, 564, 570, 577, 580, 584–5, 631, 635–6
Dupré, J. 267, 277–8, 326
Dworkin, R. 529, 534, 646
- Earman, J. 276, 307, 553, 579, 639
Ebbs, G. 155
Eberhardt, F. 276
Edelman, G. M. 159
Edinow, W. 153
Edmonds, D. 153
Eells, E. 277, 325–6
Einstein, A. 50–1, 58, 75, 152, 166, 256, 276, 297, 299, 302–4, 306, 309, 340, 553, 559–60
Ellis, B. 268, 277
Elsamahi, M. 278
Enc, B. 435
Enoch, D. 484, 494, 496
Epicurus 285
Estlund, D. 541
Euclid 18, 20, 27, 50, 153, 286, 288, 299, 300, 595
Evans, G. 121, 133, 209, 577, 585
Everett, H. 305
Ewing, A. C. 558
- Fair, D. 322
Falk, W. D. 484–5
Fantl, J. 438, 443
Faraday, M. 267
Farkas, K. 650
Feigl, H. 50, 154, 583, 629
Feinberg, B. 149
Feldman, R. 436, 440
Feyerabend, P. 260–1
Feynman, R. 121, 308
Field, H. 163, 189–90, 192–3, 599
Fischer, J. M. 426–8
- Fish, W. 464, 472
Fitelson, B. 276
Floyd, J. 152
Fodor, J. 134, 159, 261, 277–8, 364, 367, 373–4, 570, 584, 635
Foot, P. 101–2, 484, 584, 635
Forbes, G. 252
Forster, E. M. 150
Foster, J. 153, 159, 577, 581
Fraassen, B. 266, 269, 277–8, 585
Frances, B. 169, 436
Frank, P. 50, 58, 152
Frankfurt, H. 416, 427, 642
Freeman, A. 554
Frege, G. xi, 13, 19, 23, 25–31, 33, 39, 41, 46, 49, 51, 65, 72–3, 82, 118, 121, 124, 149, 151–2, 158, 163, 172, 174–6, 182, 185–8, 203, 338, 556, 561–2, 564, 569–70, 574, 580–4, 589–91, 594, 605, 611, 618, 620–1, 631, 635–7
French, S. 272
Friedman, M. 153, 278, 579, 634
Frigg, R. 277
Fry, R. 150
Fumerton, R. 435
- Gaskin, R. 62, 149, 164–5, 561, 564, 580
Gaynesford, M. 158
Geach, P. 339, 346, 570
Gendler, T. 555, 644
George, A. 27
Gibbard, A. 56, 106, 153, 585, 594, 645
Giere, R. 277
Gillett, C. 277
Ginet, C. 424
Ginsborg, H. 251
Glennan, S. S. 324
Glock, H. J. 575, 578, 581, 636
Glüer, K. 251–2
Glymour, C. 276
Gödel, K. 50, 120, 151, 163, 181–6, 197, 583, 602, 636
Godfrey-Smith, P. 332
Goff, P. 554
Goldbach, C. 178
Goldfarb, W. 249, 251
Goldman, A. 132, 440–1

- Goldstein, L. 154, 633
Goldstein, S. 309
Goodman, N. 37, 259, 276, 565–6, 577, 583–4, 635–6
Greco, J. 442, 446–7
Greene, B. 308
Grice, H. P. 312, 470–2, 635
Grice, P. 83–5, 94–8, 156, 164, 206, 215–19, 221, 223, 225, 637
Griffin, N. 561
Grimm, S. 443, 448
Gutting, G. 576–7, 581
- Habermas, J. 157
Hacker, P. M. S. xiii, xiv, 83, 156, 561–2, 564, 579–80, 633, 635
Hacking, I. 271, 275–8, 638
Haddock, A. 251–2
Hahn, H. 50, 152
Hahn, L. E. 157
Hájek, A. 276
Hale, B. 163, 186–7
Hall, N. 332, 640
Hampshire, S. 83, 156
Hanson N. R. 261, 277
Hardin, C. L. 271, 278
Harding, S. G. 277
Hare, R. M. 83, 99–103, 105–7, 156, 169–70, 583–4, 594, 644
Harker, D. 278
Harman, G. 158, 278, 579, 594
Harnish, R. M. 85
Hart, W. D. 83, 156, 584, 635
Hartmann, S. 276–7
Hattiangadi, A. 251
Hawthorne, J. 438–9, 443, 555, 644
Hazlett, A. 169, 438–40, 443
Heathcote, A. 317–18
Heck, R. 27
Hegel, G. x, 4, 9, 11, 149, 552, 572, 575–7
Heidegger, M. x, 53, 153, 572, 583
Heil, J. 330
Hellman, G. 151, 192–4, 198
Helmholtz, H. 267, 581
Hempel, C. 57, 87, 129, 153, 258, 265, 267, 276, 278
- Herman, B. 496
Hesse, M. 263, 277
Hilbert, D. 51, 151–2, 163, 179–83, 301, 602, 636
Hill, T. 496
Hintikka, J. 570, 629
Hintikka, M. B. 277
Hinton, M. 462, 472
Hirsch, E. 561, 565
Hobart, R. 425
Hobbes, T. 104, 170, 343, 481, 485, 567
Hooker, B. 508
Hooker, C. A. 278
Horton, J. 547
Hume, D. 58, 106–7, 163, 166–7, 169–70, 172–4, 187–8, 197–8, 240, 256, 265, 313–18, 320–2, 326, 343, 423, 457, 472, 484, 533, 547, 572–3, 612, 614, 622, 640
Huoranszki, F. 169, 420, 428–30
Hurley, P. 518
Husserl, E. 561, 581–2
Huxley, T. H. 128
Huygens, C. 259
Hylton, P. 11, 24, 71, 152, 155, 561, 635
- Irzik, G. 277
- Jackson, F. 106, 133, 168, 355–6, 358–60, 373, 375, 562
Jaynes, E. T. 260
Joachim, H. 3
Johnston, M. 168–9, 375, 412, 555, 559, 642
- Kant, I. x, 4, 8–9, 15, 20, 25, 50, 73, 91, 94, 114, 153, 163, 172–4, 176, 178–83, 186, 197, 240, 256–7, 481–2, 485–6, 495–6, 506, 531, 533, 562, 567, 573, 576–7, 591, 603, 618, 635, 637, 644
Kapitan, T. 422
Kaplan, D. 158, 226, 570
Kasrils, R. 149
Kearns, S. 494
Kelly, T. 448
Kenny, A. 152, 155, 631, 633
Kepler, J. 75, 287

- Keynes, J. M. 148, 150, 260
 Kieseppä, I. A. 277
 Kim, J. 328
 Kitcher, P. 271, 278
 Klosko, G. 532, 536, 547
 Kneale, M. 172
 Kneale, W. 172, 558
 Knowles, D. 278
 Kolmogorov, A. 259
 Kornblith, H. 436, 443, 447
 Korsgaard, C. 485, 489, 495–6
 Kriegel, U. 556
 Kripke, S. 118–23, 135–6, 158, 164–5,
 167, 230–9, 241–2, 244, 247, 251–2,
 351–2, 358, 360, 369–70, 435, 556–7,
 562, 570, 574, 579, 584–5, 609, 611,
 619, 633, 635, 637–8
 Kronecker, L. 21
 Kuhn, T. 59, 258, 260–2, 275–6, 584,
 635, 638
 Kusch, M. 439
 Kvanvig, J. 435, 444–5, 447

 Ladyman, J. 165, 272, 275–8, 306,
 554, 638
 Lakatos, I. 263, 276
 Lange, M. 271, 277–8, 638
 Lasnik, H. 227
 Laudan, L. 271, 276–7
 Leibniz, G. W. 10, 18, 91, 176–7, 198,
 259, 289, 300, 306, 337–9, 344, 458–9,
 558, 566, 573, 580–1, 602, 609
 Leiter, B. xiv, 563, 581
 Leng, M. 163, 192
 Leplin, J. 277
 Lepore, E. 157, 226
 Leslie, S. J. 391, 412
 Leucippus 285
 Levinas, E. 576
 Levine, J. 135
 Lewis, C. I. 609
 Lewis, D. 132, 155, 167, 260, 265, 277,
 309–10, 312, 316, 319–23, 328–30,
 332, 340, 344–5, 348, 355, 360–1, 365,
 370–3, 375–6, 404, 412–13, 437–8, 535,
 557–9, 563, 570–1, 581, 584–5, 592,
 595, 609–10, 615, 635, 639–40, 642
 Lewis, P. 271, 278
 Lipton, P. 278
 Loar, B. 556
 Locke, J. 9, 82, 89, 104, 170, 231, 267,
 338, 349, 351–2, 379, 394, 401, 404,
 411, 472, 481, 485, 532–3, 567, 572,
 574, 589, 609, 622–3
 Lockwood, M. 136
 Loewer, B. 165–6, 252, 277, 308–9, 553
 Logue, H. 376, 472
 Lorentz, H. 276, 297–8, 304, 308
 Loschmidt, J. 293
 Lowe, E. J. 337–8, 340, 342, 344–5, 347,
 350–3, 424, 640
 Luckhardt, C. G. 152
 Ludlow, P. 227, 637
 Ludwig, K. 157
 Lynch, M. 447
 Lyons, T. 278

 Mabe, M. 552
 McDowell, J. 12, 106, 127, 133, 236,
 251, 463, 472, 564, 577, 585, 594, 645
 MacFarlane, J. 438
 McGinn, C. 126, 135, 159, 237–8, 252,
 562, 580
 McGrath, M. 438, 443
 Mach, E. 50–1, 152, 256, 267–8, 290,
 307, 629
 Machamer, P. 324
 Mackie, J. L. 157, 277, 316, 412, 455,
 472, 489, 495, 577
 McMullin, E. 268
 Macpherson, F. 555
 McTaggart, J. 3–4, 19, 148–9, 624
 Maddy, P. 190
 Magee, B. 23, 90, 153, 156
 Magnus, P. D. 271, 278
 Malament, D. 189
 Marcus, R. B. 158, 570
 Markovitz, J. 496
 Marr, D. 134
 Maslen, C. 330
 Matar, A. xiv, 563
 Maudlin, T. 306, 308–9, 579, 639
 Maxwell, J. 50, 166, 267, 296–8, 302, 308
 Mayo, D. 276

- Meacham, C. J. G. 276
Meijs, W. 276
Meinong, A. 51, 151, 570, 581
Mele, A. R. 428
Melia, J. 191–2
Mellor, D. H. 277, 328
Menger, K. 50–1
Menzies, P. 316–17, 324, 329, 640
Meyer, E. 277
Michelson, A. 297
Miklós, A. 514
Miklósi, Z. 514
Mill, J. S. 15, 51, 170, 256, 265, 267, 548, 611, 614
Miller, A. 251, 637, 645
Miller, D. 548
Millikan, R. 449
Minkowski, H. 297–9, 304, 309, 340
Montague, M. 557
Monton, B. 278
Moore, G. E. xi, 3, 8–17, 29, 37, 40, 54, 56, 63, 71, 84, 87, 88, 99, 102–3, 106, 117, 124, 148–50, 154, 169, 435, 484, 555–6, 561, 563–4, 568–70, 573–4, 580–3, 590, 594, 604, 622, 628, 632, 635, 644
Moran, R. 374
Morley, E. W. 297
Morris, M. 152
Mulgan, T. 504, 506
Mulhall, S. 529
Murphy, L. 507, 513–15
Musgrave, A. 276

Nagel, E. 135, 265
Nagel, J. 443
Nagel, T. 57, 130–1, 133, 135, 158–9, 168, 265, 267, 358, 442, 456, 472, 483–4, 545–6, 584, 642
Nerlich, G. 553
Neurath, O. 50, 53, 56–7, 76, 152–4
Newton, I. 18, 50, 75, 166, 176–7, 189–90, 198, 256–8, 263–4, 267, 286–92, 294–302, 306–8
Nichols, S. 484, 580
Niiniluoto, I. 277
Noordhof, P. 320
Nowell-Smith, P. H. 156
Nozick, R. 105, 170, 435, 440, 584, 642, 645

O'Connor, T. 329, 423
Oliver, A. 277
Olson, E. 346, 412, 642
Olstrom, E. 535
O'Neill, M. 541
Ortiz-Milan, G. xii
O'Shaughnessy, B. 496
Otsuka, M. 532
Owens, D. 277

Papineau, D. 268, 270, 275–8, 448–9
Parfit, D. 379, 382, 386, 412, 484–5, 487–9, 494–7, 543, 555, 567–8, 573, 577, 585, 594, 642, 645
Parker, W. 278
Pascal, B. 70, 259
Pasteur, L. 267
Pauli, W. 303
Pautz, A. 375
Peacocke, C. 133, 466, 472, 577
Peano, G. xi, 19–20, 23, 25, 51, 150, 174, 183, 187, 192–3, 195, 581
Pearl, J. 326
Pears, D. 83, 156
Peirce, C. S. 256
Penrose, R. 159, 308
Pereboom, D. 423, 428
Perez, D. xii
Perry, J. 158, 221–2, 386
Peterson, M. 510–13
Pettit, P. 106, 540
Pietroski, P. 227, 277
Pigliucci, M. 276
Pitcher, G. 465, 472
Pitt, D. 556
Place, U. T. 131
Planck, M. 152
Plantinga, A. 440, 571, 581, 584
Plato 12, 65, 108, 169–70, 257, 614, 637
Platts, M. 160
Pogge, T. 543–5
Poincaré, H. 51, 152, 255, 268, 595

- Popper, K. 23, 57–9, 153–4, 255–8, 261, 263, 270, 276, 583–4, 638
- Portmore, D. 508, 517–20
- Price, H. 316, 318, 558, 573, 629, 644
- Prichard, H. A. 16, 484
- Prinz, J. 553–4
- Pritchard, D. 442, 445
- Pryor, J. 435
- Psillos, S. 268, 270–1, 277–8
- Ptolemy 75, 286
- Pufendorf, S. 481, 485
- Putnam, H. 118–21, 123–6, 132, 158, 167, 188–9, 192, 194, 198, 264, 268, 272, 351–2, 472, 556, 579, 584–5, 635
- Quine, W. 23, 71–82, 108, 110, 113, 115, 122–3, 129, 143, 146, 154–5, 157–8, 163–4, 184, 186, 188–91, 197, 204–5, 255, 264, 272, 277, 316, 370, 552, 556–7, 561, 563, 570, 573, 577, 583–4, 591–2, 597–8, 600, 617, 622, 635–7
- Quinton, A. 403, 413
- Quong, J. 529
- Radford, C. 439–40
- Railton, P. 106, 484, 495–6, 594, 645
- Ramsey, F. 23, 61, 132, 154, 260, 265, 277, 583
- Ravizza, M. 427–8
- Rawls, J. 105–6, 157, 170–1, 484, 505, 514, 528–30, 534–44, 547–8, 557, 570, 584–5, 605–6
- Raz, J. 484–6, 531, 537
- Recanati, F. 85, 214, 221–5
- Redhead, M. L. G. 553
- Reichenbach, H. 56, 154, 256, 276, 322–3, 553, 583, 629
- Rescher, N. 277
- Resnik, M. D. 194
- Rey, G. 277
- Richardson, A. 153, 579
- Robb, D. 330
- Roberts, J. 277
- Robinson, H. 131, 159, 358, 375, 459, 461, 464, 472, 555
- Roessler, J. 472
- Rosen, G. 571, 579, 581, 599
- Rosenberg, A. 271, 275, 277–8
- Ross, W. D. 16, 102, 306, 484
- Rousseau, J. J. 104, 532, 535
- Ruben, D. 278
- Russell, B. xi, 3–4, 8–13, 18–26, 30–42, 46, 48–9, 51–2, 61, 65, 70–2, 78, 82–3, 88–91, 109, 117–18, 121, 124, 127, 136–7, 148–52, 154–6, 158, 160, 163, 166, 169, 174–6, 179, 182, 185, 187–8, 313–16, 324, 556, 561, 563–5, 569–70, 573–4, 580–4, 590–1, 596–7, 605, 607, 611, 614–5, 619, 628–9, 631–2, 634–7
- Ryle, G. 83–4, 87–9, 98, 130–2, 149, 156, 357–9, 361, 373–5, 440, 583–4, 593, 635, 641
- Rysiew, P. 438–9
- Saint Augustine 66
- Salmon, M. H. 276
- Salmon, N. 158
- Salmon, W. 167, 276–8, 322–3
- Sangiovanni, A. 544
- Sartwell, C. 440
- Saunders, S. 309, 553
- Savage, L. J. 260
- Scanlon, T. 105, 484, 489, 494–6, 543, 567, 585, 642
- Schaffer, J. 319, 322, 329–31, 438, 580
- Scheffler, S. 514, 518, 548
- Schilpp, P. A. 150, 566
- Schlick, M. x, xii, xiii, 50, 52, 55, 57, 71, 148, 153–4, 553, 583
- Schrödinger, E. 301–5
- Schroeder, M. 484, 495
- Searle, J. xiii, xiv, 85, 134–5, 159, 556, 584–5, 623, 634–5, 637
- Sellars, W. 573, 583–4
- Sen, A. 529
- Shafer-Landau, R. 484, 494
- Shapere, D. 277
- Shapiro, S. 163, 190, 194–6
- Shiffrin, S. 542
- Shoemaker, S. 374, 386, 398, 401, 404, 413, 584–5
- Sider, T. 448, 559, 565–6, 579, 585
- Siderits, M. 411

- Sidgwick, H. 17, 102–3, 150, 484
Simmons, A. J. 531, 533, 535
Singer, P. 494, 584
Skinner, B. F. 129–30, 132
Sklar, L. 277–8, 307, 584, 638
Smart, J. J. C. 131, 355, 570
Smith, A. P. 148
Smith, B. 164, 227
Smith, D. 472
Smith, M. 484, 495
Snowdon, P. 169, 376, 412, 472, 555
Soames, S. 157–8, 551–2, 636
Sobel, D. 507
Sokal, A. 553, 585
Sosa, E. 277, 435, 438, 440, 442, 446–7
Sperber, D. 221, 227
Spinoza, B. 143, 395
Stainton, R. 211, 227
Stalnaker, R. 360, 365–6, 376–7, 570
Stanley, J. 222, 225–6, 359, 438–9, 443
Star, D. 494
Steiner, H. 532
Steiner, M. 197
Stevenson, C. L. 56, 99, 102–3, 153, 635
Stich, S. 375, 443
Stoljar, D. 136, 168, 356, 359, 362, 375
Stout, G. F. 558, 581
Strachey, L. 150
Strawson, G. 135–6, 159, 412, 417, 555–6, 590
Strawson, P. F. 83–5, 89–93, 99, 156, 240, 557, 570, 577, 584, 623, 634–5, 642
Stroud, B. 435
Stroud, S. 519
Sturgeon, N. 106
Suppe, F. 277
Suppes, P. 108, 325
Swift, A. 529
Swinburne, R. 276, 413, 430
Szabo, Z. 226

Tännsjö, T. 516
Tarski, A. 58, 71, 111–13, 138–43, 157–8, 571, 583, 620–1, 637
Thomson, J. J. 494
Tiberius, V. 484

Tobin, E. 277
Tooley, M. 266, 277
Travis, C. 472
Treanor, N. 448
Turing, A. 24 *see also* Turing Machine
Tye, M. 133, 375, 579

Uebel, T. 153, 579
Unger, P. 440
Urbach, P. 276
Urmson, J. O. 156
Valberg, J. 412, 555
Van Inwagen, P. 332–3, 346, 422, 425, 558, 570, 585, 642
Vassalo, N. 157
Verheggen, C. 243, 251–2
Vihvelin, K. 429
Vogel, J. 435
von Mises, R. 50, 152

Waismann, F. 50, 153
Waldron, J. 534–6
Wallace, D. 309
Wallace, R. J. 427, 484
Walton, A. 545
Warnock, G. 83, 156
Wedgwood, R. 449, 484
Weierstrass, K. 18–19, 176–7
Weinberg, J. 443
Weinberg, S. 412
Wellman, C. 535, 547
Whewell, W. 256–7
White, R. 436
Whitehead, A. 18, 22–4, 40, 51–2, 108, 148, 152, 176, 564, 582
Whiting, D. 251
Wiggins, D. 339, 346, 385, 570, 577, 585, 623
Wikforss, A. 251–2
Wilkes, K. 412
Williams, A. 540
Williams, B. 157, 385, 439–40, 484–5, 494–6, 502–4, 584–5, 642
Williams, D. C. 558
Williamson, J. 316, 318, 324, 332
Williamson, T. 359, 442, 541, 552, 562, 570–2, 585, 636

- Wilson, D. 221
Wilson, T. 374
Winsberg, E. 277
Wittgenstein, L. xi, xiii, 12, 24, 39–49,
50–2, 57–8, 61–70, 83, 88, 109, 130–1,
148, 152–7, 164–5, 230–1, 235–41, 244,
246, 249, 251, 253–4, 369, 440, 556, 561,
563–4, 566, 569–70, 573–4, 579–80,
582–3, 585, 590, 596, 599, 601–2, 605,
608, 616, 628–9, 632–3, 635, 637
Wollheim, R. 412
Woods, M. 577
Woodward, J. 326–7
Woolf, V. 150
Woolhouse, R. 276
Woozley, A. D. 156
Worrall, J. 271–2, 278
Wright, C. 163, 186–7, 234, 251,
374, 580
Yablo, S. 191, 331, 412, 561
Zagzebski, L. 446–7
Zahar, E. 259, 276
Zermelo, E. 176, 183, 186, 293, 593
Zimmerman, D. 558, 573–4

Subject Index

- a posteriori materialism 360, 372, 375
a priori scrutability thesis 567
Absolute, the 4, 6, 7, 53
accidental regularities 265, 315
Agony Argument 487–8, 496
analysis 590
 a priori conceptual analysis 123
 analysis of language xi, 14–15, 32–4,
 53, 65, 83, 88–9, 91, 569–70, 590
 conceptions of conceptual
 analysis 569
 conceptual analysis (as analytical
 philosophy) 391
 conceptual analysis (limits of) 454,
 553
 conceptual analysis of set theory 24
 counterfactual analysis 318, 321–2,
 332
 decompositional analysis 11, 563,
 573, 590
 logical analysis xi, 23, 27, 57, 63
 mathematical analysis 176, 178–9
analytic movement xi, xii, xiii, xiv, 148,
 561–3, 568–9, 580, 584
analytic propositions 46, 173
analytic truths 119, 122, 174, 187, 591
analytic/synthetic distinction 73, 184,
 186–8
animalism 403
anomalism of the mental 109
anti-Cartesianism 633
anti-naturalism 17, 99, 486, 568
anti-realism 117, 158, 251, 264, 267–9,
 302, 448
anti-reductionism 382 *see also*
 non-reductionism
argument from illusion 458–60, 472,
 622
Aristotelian ethics 15, 101
assertibility-conditions 233, 236
associational account of justice 544
associative learning 314
atomistic realism 8, 77
axiom of infinity 23
axiomatic coherence 195
axiomatic theory of meaning 109–10

Bayes' Theorem 260
Bayesianism 260
beauty 16, 17, 150
behaviorism 87, 601, 641
behaviorist psychology 129–32
belief theory of perception 465
biconditionals 110, 112, 115, 592
biological determinism 418, 420–1
Bloomsbury Group 150
Bohmian mechanics 305, 308
bound variable 78, 142, 224
Bradley's regress 6, 13, 564, 580
Buddhism 378

calculus 18, 33, 141, 150, 174, 176–7,
 286, 590, 603, 615
Canberra Plan 592
Cantor's Paradox 179–80
cardinality 177–9, 592
Cartesianism 135, 264, 357–8, 373, 379,
 556, 599, 616

- causation 166–7, 264, 273–4, 312–32, 401, 423, 478, 592, 640
 - agent causation 424–5
 - causal relevance 329–31
 - causation by disconnection 331–2
 - constant conjunction 313–14
 - folk theory of causation 324, 592
 - regularity theory of causation 167, 315–16, 318
- class 20–3, 32, 59, 175, 326, 337, 593
- Closure Principle 433
- commonsense intuitions 385–7, 390–1, 393, 397–401, 405, 428, 443–4, 562, 603–4
- communicative intentions 95
- communitarians 529
- compatibilism 169, 418, 422, 425–6, 428–9, 643
- compositional change 342–5
- compositional theory of truth 111, 207, 209
- compositionality 109, 111, 209
- computation 132, 134, 230–1, 247–8, 274, 373, 601
- conceptual schemes 115
- conceptualism 353
- connectionism 134
- conscious states 7, 16, 93, 150, 355, 362–3, 374, 556, 594
- consent theory 530–5, 547
- consequence argument 169, 421–3, 426, 429
- consequentialism 171, 474, 477, 500–20, 567
- conservativeness claim (mathematics) 190
- conserved quantity theory 323–4
- constructive conceptual analysis 563
- constructive empiricism 269
- contextualism 225, 227, 437–8, 595, 643
- contractualism 105, 567
- contrastive relation 330
- conversational implicature 96, 98, 219, 223, 225, 438
- cooperative principle 97
- cosmopolitanism 171, 543–4, 546, 548
- counterfactual conditionals 237, 247
- counterfactual dependence 318–20, 328, 332, 639
- covert behavior 129
- criterion of demarcation 58, 256–7, 276
- cultural relativism 115
- decision theory 108, 274–5
- definite descriptions 31–3, 78, 121, 226, 365, 611, 619
- demandingness objection (moral) 171, 500–20
- denoting concepts 32, 121, 597, 616–17, 631
- descriptive metaphysics 91–2, 557
- descriptive phrases 31, 35, 597
- descriptive psychology 168, 367–70, 374–5
- descriptivism 119–21
- determinism 169, 259, 290, 295, 307, 417, 418–30, 643
- dialectical conception of justification 440
- direct realism 123, 169, 273, 644
- direct reference 118–20, 158
- disguised descriptions 33, 590
- disjunctivism 369, 376, 462–5, 470–2, 644
- dispositions,
 - behavioral dispositions 130, 132, 157, 232, 421, 641
 - causal dispositions 250
 - dispositions and normativity 237–8, 241, 245, 247–8, 480–4, 486
 - innate dispositions 133
 - mental dispositions 130, 618
- division of (moral) labor 513–14
- dualism 131, 136, 157, 159–60, 355, 357–8, 361, 365, 368, 374, 570, 581, 642
- Duhem problem 262, 277
- Dutch Book Argument 259
- eliminative materialism 134
- emotivism 56, 100, 106, 169–70, 593, 597

- empiricism 4, 50–2, 73, 76, 152, 173,
184, 255–6, 262, 268–9, 317, 570,
615, 632
- endurance 380, 412
- epiphenomenal consciousness 128, 136
- epiphenomenal properties 331
- epistemic expressivism 449
- Equivalence Thesis 138–9
- essence 69, 118, 167, 328, 337, 347,
348–53, 557
- essentialism 122, 343, 349, 352–3, 598
- ethical judgments 54–5, 153, 612
- European Society for Analytic
Philosophy xii, 578
- evidentialism 440–1
- evidentialist conception of
justification 440
- exclusion problem 330–1
- experiential streams 554
- experiential transparency 357, 373,
555
- experiential unity 6–7, 554
- experimental philosophy 562
- explanatory gap 135
- extensionality 13, 27, 111–12, 598
- externalism 119, 126–7, 198, 480–1,
495–6, 599, 626
- fallibilism 258
- falsifiability 58
- family resemblance 69, 70, 580
- fatalism 418–20
- fictionalism 163, 186, 189, 191–2, 195,
599
- first-order logic 78
- folk psychology 273
- foundational psychology 370, 373
- free enrichment 224–6
- free will 169, 255, 273, 415–30
- functionalism 130–4, 159, 324, 472, 601,
641
- general relativity 50, 259, 270, 299, 304,
340, 553
- generative grammar 205
- grammar (as a guide to logical
form) 34–5, 89, 590
- grammar 203, 238, 462, 561, 601
Quine v. Chomsky 204
- grammatical forms 63–4, 88, 164, 205,
207
- Great Expansion, the 106, 169
- hallucination 86, 457, 459–60, 463–4,
472, 554
- hermeneutics 57
- historicism xiii
- holism 7, 190
- human flourishing 15, 101, 104
- Hume's Principle 187–8
- hypothetico-deductivism 257–8
- identity 28, 131, 336–7, 339, 602
criteria of identity 338, 340, 346–7,
350, 352–3
diachronic identity 342–3, 393, 398
identity of form 45
identity statements 121–2
identity theory of truth 12
numerical identity 342, 380
personal identity 93, 168, 338, 378,
398, 400–1, 405–9, 424
qualitative identity 342
synchronic identity 342
transworld identity 348–9
- ignorance hypothesis 361
- illocutionary act 86, 623
- incompatibilism 418, 420–2, 428–9
- incompleteness theorems 151, 181,
183, 602
- indefinite reading 224
- indeterminism 264, 296, 423
- indirect realism 273, 644
- indispensability argument 188
- induction 58–9, 166, 257–8, 269–71
- infinitary mathematics 163, 176, 178–82
- infinitesimals 19, 177
- intensional entities 72
- intentional states 230, 363, 367
- intentionalism 465–6, 472
- intentionality 236, 356, 362–75, 415,
555–6, 603
- internalism 127, 480–1, 599
- intersubstitutability 28, 598

- intuition (Kantian) 20, 25, 173, 176, 178–9, 603
- intuitionism (mathematical) 163, 178, 182, 604
- intuitionism (moral) 17, 54, 56, 99, 106, 604
- intuitions (about personal identity) 385–91, 397
- irrational numbers 23, 177

- justice 105, 157, 170–1, 528–30, 534–8, 645–6
 - principles of justice 105, 170, 528–30, 536, 538–47
 - relational accounts of justice 544
- Knowledge Argument 133, 358–63, 365
- knowledge by acquaintance 35, 86, 591, 611

- language xiii, 27, 31–5, 41, 43, 45–7, 49, 53, 58–9, 63–8, 72, 77, 79–85, 94, 96, 98, 109–18, 138–42, 152, 154, 155, 164, 201–15, 230, 233–5, 243, 365–7, 437, 467, 556, 563, 566, 570, 580, 600–1, 603, 608, 612, 616, 621–3
 - artificial languages 52, 140
 - language-game 67–8, 70, 91, 98, 236, 601
 - metalanguage 110, 138, 154
 - ordinary language xi, 34–5, 45, 63, 83–4, 87, 89, 91, 94, 98, 109, 118–19, 156–7, 167, 169, 563, 567, 569, 572–3, 590, 605–6, 612–13, 622, 633–4
 - private language 165, 230, 616
 - public language 243, 556, 608
 - simple languages 140–1
- Leibniz's Law 337, 338, 339, 344, 458, 459
- Liar Paradox 432
- liberal egalitarianism 528, 538
- libertarianism 170, 423–5, 430, 643, 645
- linguistic community 80, 158, 233–4
- linguistic framework 59, 77
- linguistic idealism 564, 580

- logic xiii, 20–7, 31, 34, 40–1, 45–9, 51–3, 58, 65, 70–8, 91, 150, 172, 174, 176, 179, 182, 187–8, 190, 196–7, 240–1, 255–8, 262, 336–7, 572–3, 594, 600, 603–7, 609, 625
- logical positivism 17, 50, 58, 262, 613, 615, 632
- logicism 24, 27, 150–1, 604, 607, 635, 637
- luck egalitarianism 529

- Manipulation Argument 420, 426, 428
- many worlds interpretation (of quantum theory) 305, 348, 553
- material coincidence 346
- meaning,
 - Chomsky and meaning 164
 - communal meaning 235
 - compositional meaning 222
 - conventional meaning 52, 96
 - Davidson and meaning 109–16, 157
 - E-language 206
 - extensional account of meaning 27, 112
 - factuality and meaning 234
 - falsity and meaning 91
 - the “golden triangle” and meaning 562
 - Grice and meaning 94–6, 98, 215–17
 - I-language 205
 - implicit meaning 219
 - indefinite meaning 224
 - indeterminacy of translation 79–81, 556
 - irreducibility of meaning 165
 - literal meaning 223
 - logic and meaning 34
 - mathematics and meaning 181
 - mental image 231
 - natural meaning 94–5
 - normativity of meaning 242
 - positivists and meaning 52–8
 - primary intension 159
 - the problem of meaning 202, 230
 - Putnam and meaning 123–6
 - Quine and meaning 79–81
 - reference and sense 28–9, 31, 33, 118, 120, 562, 621
 - synonymy 73

- Tarski and meaning 138–42
theories of meaning xii, 555
utterance 85, 95, 164, 209–10, 212–13
Wittgenstein and meaning 44–8,
66–7, 237–9, 601
measurement problem 302, 304, 639
Meno question 447
mental states 116, 125, 129–34, 232,
272, 323–4, 329–30, 372–5, 466, 472,
480, 596, 601, 603, 608, 618
mereological essentialism 343
mereological nihilism 558
mereological universalism 558
metaethics 14, 15, 106
metametaphysics 336, 561, 565
metaphysical realism 116, 565
methodology (philosophical) x, 553,
563, 571
methodology (scientific) 57, 165,
256–7, 261, 267–8
mind-independent reality 8, 11, 302,
314, 353
Minkowski space-time 297–9, 304, 309
modal realism 558, 609–10
modal structuralism 186, 194–5
model-theoretic theory of truth 58
moral duty 102, 150, 477, 531, 534,
542–3, 546
moral evaluations 56, 101
moral obligation 102, 242, 410, 534
moral rationalism 517–20
moral realism 106, 157, 568
moral truth 54, 568, 598, 604
multidimensional
 consequentialism 171, 510–13
multiple realizability 130, 330, 601
Naïve Semantic Picture 164, 207, 209,
212, 214
natural duty 530, 534–5, 537, 546
natural necessity 265
naturalism 15, 17, 73, 76, 82, 128,
134–6, 188, 449, 486, 553–4, 579
naturalistic dualism 136
naturalistic ethics 99, 568
naturalistic fallacy 15, 16
naturalized epistemology 75
neo-Hegelian idealism 3, 19
neo-Humanism 167
neo-Kantianism 50, 71, 485, 496, 577,
634
neo-Lockeanism 403–4
neo-logicism 163, 186, 187–8, 190, 195,
197–8, 564, 580
neo-Mooreanism 435
Newtonian mechanics 257, 263,
286–92, 295–300, 306–8
Nicod's criterion 258–9
nominalism 189, 191–3, 196–7, 611–12,
617
noncognitivism 17, 56, 101, 593, 597
nonconceptual content 133
nonlinguistic behavior 231–2
nonreductionism 131, 379 *see also* anti-
 reductionism
nonreductive physicalism 130–1, 134
nonrigid designator 121
nonsentential assertions 227
normative externalists 480
normativity 170, 242, 250–1, 476–84,
486, 488–96, 612
noumena 4

objectivism 54, 106
objectivity 92, 580
Okham's Razor 262
one-to-one correspondence 20, 143,
177, 592
out-of-body experiences 388

panpsychism 554, 570
paradox of analysis 88
paradox of decrease 345
paradox of increase 345
parity principle 187
particularity objection 535
perception 36, 86, 169, 273, 369, 433,
453–6, 459, 461–3, 465, 467, 470–2,
554, 622, 644
perceptual concepts 454, 470–1
perdurantism 380, 412, 559
perlocutionary act 86
persistence 336, 339–44, 346–7, 379–81,
384–5, 387, 390, 393, 403, 405

- persistence conditions 340–1, 346–7, 380–1, 385, 387, 395, 399–401, 403
- personal responsibility 409, 417–18, 424–5, 427, 529, 534, 541, 544
- phenomenal intentionality [PI]
 - program 555–6
- Phenomenal Principle 459
- philosophical anarchists 531
- picture theory 44
- Platonic atomism 10, 11, 21, 564
- Platonic forms 46–7
- Platonism 17, 163, 186, 191, 194, 196, 604, 614
- Platonist view of mathematical objects 184–6
- political obligation 105, 531–2, 534, 546
- political values 261, 276, 528–9
- possible worlds 119, 122, 159–60, 291, 321–2, 328, 348–53, 558, 571, 589–90, 596, 599, 609–10, 615, 619
- practical rationality 170, 256, 485, 490
- practical reason 9, 170, 474–8, 485–6, 488, 490–4, 496
- pragmatic encroachment 438, 443
- predicate logic 31, 78, 174, 176, 188, 603
- presentism 344–5, 560, 600, 616
- primitive property 15
- Principal Principle 260
- Principle of Charity 113, 115–16, 369
- Principle of Indifference 260
- Principle of the Identity of Indiscernibles 289, 602
- privileged access 374
- probabilism 259
- probability 257–60, 276, 293–6, 301–8, 324–6, 604, 637
- problem of temporary intrinsics 340, 344–5
- process theories 167, 318, 332
- promiscuous realism 267
- proposition 9, 617, 622
 - acquaintance and propositions 35–7
 - Ayer and propositions 53–5
 - contextualism and propositions 225
 - empirical propositions 76
 - Hume and propositions 172–3
 - meaning and propositions 82
 - mind-independent propositions 34
 - minimal propositions 220–1, 226
 - modal propositions 353
 - objective propositions 13
 - Platonic conception of propositions 10, 12, 89
 - proposition radical 219, 223
 - propositions as terms 10
 - Syntactic Correlation Thesis 216–18
 - truth and propositions 11–12, 19, 90, 419–20, 448
 - truth-theory and propositions 209
 - Wittgenstein and propositions 42–9, 63, 65, 70
- propositional attitudes 116, 126, 440, 617, 619
- propositional content 219, 466–7
- propositional functions 19, 31, 37, 151, 604
- propositional knowledge 359
- propositional representational content 466–7
- propositional unity 564–5, 632
- protocol sentences 53, 57, 617
- psychological continuity 398, 404–7, 555, 642
- psychological determinism 421, 425, 430
- psychological externalism 119
- psychological reductionism 404–7
- psychologism 20, 25, 556, 618
- psychophysical laws 131, 136
- qualia 133, 168, 355
- quantifier variance thesis 561, 565
- quantifiers 19, 606, 618, 620
- quantum theory 165–6, 273, 288, 300, 302, 304–6, 308–9, 553, 638–9, 643
- radical interpretation 112–13
- rational action 106–7, 170, 476–7, 484–5, 490
- rational numbers 177
- rationality xiii, 106, 114, 260–1, 485, 489, 577
- Raven Paradox 259, 276

- real numbers 23, 177
- realism 158, 267, 271, 278, 448, 611, 614, 645
- reasonable pluralism 529
- recursive clause 141
- recursiveness 109
- reductionism 57–8, 130–1, 262, 364, 404, 618
- referring expressions 28, 31–2, 88, 121, 579, 596, 598, 603, 607, 620
- regulative control 426
- relationism 289–91, 307, 644
- relations of ideas 172–4
- reliabilists 435, 440–2
- representationalism 134, 369, 374, 467, 554, 579
- reversibility paradox of statistical mechanics 293–4, 307
- revisionary metaphysics 91, 558–9, 592
- rigid designator 121–2, 159, 619
- rule-following 235, 238–9, 244, 249–50
- Russellian Monism 137
- Russell's Paradox 22–3, 26, 31, 39–40, 48, 163, 175, 179, 182, 187
- Russell's Theory of Descriptions 33–5, 78, 90, 118, 563, 569, 590, 597, 607, 619, 628, 634
- Schrödinger's equation 301–5
- scientific realism 81, 123, 256, 264, 267–72, 278
- second-order logic 187–8, 195, 197, 600
- self-determination 416–17, 425
- semantics 621
 - Carnap and semantics 58–9, 77
 - descriptive semantics 366–8, 374
 - indeterminacy and semantics 81–2
 - methodological semanticism 571
 - nonstandard semantics 192
 - scientific semantics 140
 - semantic explication 223
 - semantic information 222–3
 - semantic minimalism 226
 - semantical obligation 242
 - semantic-pragmatic distinction 215
 - the semantics, syntax and pragmatics trichotomy 201, 206–7, 213, 214
 - semantics/pragmatics distinction 215–18, 222–3
 - syntax and semantics 201, 531
 - two-dimensional modal semantics 136, 562
- semi-compatibilism 427–8
- sense-data 35–7, 58, 86, 151, 457–8, 460–3, 465, 472, 563, 611, 622, 644
 - see also* sense-datum; sensibilia
- sense-datum 169, 456, 458, 460–4, 472, 612, 622, 644 *see also* sense-data; sensibilia
- sensibilia 37, 86 *see also* sense-data; sense-datum
- sets 22, 77, 177–9, 185–6, 188, 196, 338, 592–3, 614, 623
- skepticism 116, 126, 157, 233–4, 270, 432, 643
- social contract 105, 531–2, 645
- social determinism 418, 420–1, 425, 430
- source externalism v. source internalism 480–1, 495
- source voluntarism 481, 495
- special composition question 558
- special relativity 166, 259, 297, 340, 559–60
- speech acts 85, 86, 90, 99–100, 623, 634–5
- statistical mechanics 166, 273, 285, 287, 291, 293, 295–7, 308
- structural realism 272
- subjectivity 133, 135
- subjects of experience 29, 92–3, 398, 555
- subject-sensitive invariantism 438
- supererogation 506
- supervenience 131, 134, 624
- syntactic structure 203–4, 206, 211
- syntactical rules 46–7, 203, 371
- synthetic propositions 173
- tautologies 47–8, 51–2
- temporal asymmetry 295

- temporal neutrality 166
temporal parts 82, 155, 340–5, 380, 382, 559–60
theological determinism 418–19
therapeutic conception of philosophy 64
thermodynamics 50, 166, 273, 292
thought experiment 168, 554–5, 562, 590
transfinite set theory 177–8, 180
transformative conceptual analysis 563
tripartite analysis of knowledge 439–42
true belief 74–6, 114, 269, 369, 432–3, 444–9
trumping pre-emption 319
truth 4, 7, 11, 12, 19, 47, 81, 90, 99, 110–12, 116, 138–40, 142, 149, 157–8, 207, 209, 236, 263, 271, 420, 432, 438–9, 449, 566–7, 571, 593, 596, 620–1, 624, 628
logical truth 40, 47, 51–2, 72, 176, 607
mathematical truth 20, 172–3, 176, 178, 183–7, 197
necessary truth 47, 52, 73, 75–6, 122–3, 135, 185, 347, 562
performative conception 90
synthetic a priori truths 47, 52, 73, 173, 176
truth as correspondence 11, 12, 90, 157, 621
truth conditions 111–12, 139–40, 222–3, 230, 233, 236, 624
truth value monism 445–7
Turing machine 132, 274
type identity thesis 130
unarticulated constituent 221
unbound variable 142, 144
uncritical semantics ('museum myth') 81
underdetermination 77, 262, 270, 277
unified science 57, 59, 129, 261, 618
universal prescriptivism 101
utilitarianism 15, 410–11, 500–20
value-based theory of content 476, 484
veil of ignorance 105
verifiability criterion 58
verification principle 53–4, 153
verificationism 53–4, 73, 153, 613
Vienna Circle x, xi, 24, 50–3, 57–61, 71, 73, 152, 154, 583, 615, 634
virtue ethics 102–4, 474
Wide Psychological View 403–4
zombie argument 135–6, 160